The Use of Verbs in Research Articles: Corpus Analysis for Scientific Writing and Translation [1]

Arianne Reimerink
Universidad de Granada

ABSTRACT
This article describes the results of a study of the use of verbs in the different sections of medical research articles. A corpus of 30 POS-tagged texts was used. The verbs were classified according to their meaning in lexical domains. Results show that the lexical domains are distributed differently in each section of the article. The comparison of these results with those obtained by López Rodríguez (2002) sheds light on the complex relationship between a research article and its abstract. The objective is to learn more about the lexical characteristics of medical English. A proposal is also made for using these results in a more practical manner, for example to enrich an electronic manual with information that could assist professionals and translators with the writing and translation of medical research articles.

KEYWORDS: academic writing, corpus linguistics, IMRAD, lexical domains, medical research articles.

Introduction

English has been the *lingua franca* for science literature, including medical science, in the Western world since the 1950s (Navarro 2001). Of all the scientific genres, research articles are the most important and prestigious communicative tool of the discourse community of scientific experts (Swales 1990; López Rodríguez 2000). Therefore, it is essential for every scientific expert who wants to be part of his/her discourse community to be able to write adequate research articles. Because of the importance of research articles and of English as the main language of the international science community, many scientists whose native language is not English prefer to publish their research results in English in international scientific journals. Professional translators are also sometimes asked to translate these highly specialized texts. Both groups of experts, scientists and translators, should be aware of the linguistic structures and recurrent lexical choices used in this genre.

The aim of this study is to examine the lexical characteristics of medical research articles in English. Research articles have been studied by genre analysts such as Swales (1990) and Nwogu (1997), who focus mostly on the macrostructure of genres and rhetorical functions or *moves* applied to different parts of research articles. Lexical characteristics of medical English have been broadly analyzed on the basis of their register in comparative studies such as Van Hoof (1998) or Pilegaard (1997) or in intralingual studies that focus on terminological issues (e.g. Tercedor and Méndez 2000; Faber and Jiménez 2002; Tercedor 2002).

This study has been carried out from the perspective of genre and focuses specifically on the analysis of lexis. Its primary objective is to examine the semantic categories of the verbs used in each of the five sections of the prototypical medical article, namely: *Abstract*, *Introduction*, *Methods*, *Results* and *Discussion*. The methodology used is based on López Rodríguez (2002) and the semantic classification of English verbs is that proposed in Faber and Mairal (1999).

It is suggested here that the results of the corpus analysis could be useful for professionals and translators with the writing and translation of research articles. An electronic manual is proposed which could make use of the type of data elicited by this study.

Theoretical background

Genre analysis by Swales (1990) and Nwogu (1997)

Two concepts studied by Swales (1990) that are essential for the corpus analysis carried out in this study are *genre* and *move*. Swales defines *genre* as "a class of communicative events, the members of which share some set of communicative purposes" (Swales 1990:58). Therefore, it can be said that the communicative purpose is the criterion that defines a genre. Discourse communities share conventional public aims and communication mechanisms, and make use of one or more genres to communicate within or between discourse communities. To be part of a discourse community one needs to know how and when to apply the conventions of each genre.

As for the concept of *move*, it refers to a text segment with a specific rhetorical value. The combination of several of these moves creates a genre which, in turn, belongs to and defines a discourse community. For example, "describing data collection procedure" is one of the moves that, together with others, defines the genre of the *research article* which belongs to all scientific discourse communities. The term *move* clearly shows that the construction of a text is an active process in which one tries to comply with the demands and restrictions imposed by a discourse community through several rhetorical steps that define and are recognized by that community. Swales (1990) only analyses the moves of the *Introduction* section of research articles while Nwogu (1997) applied this notion to all sections (see Table 1).

The above-mentioned studies by Swales and Nwogu focus on the macrostructure of research articles and the rhetorical moves expressed by the content of each section. In the present study, the aim is to focus on the lexical aspect of medical research articles and to study the possibility of combining information on macrostructure and lexis in a way useful for writers and translators of medical articles.

I will come back to Nwogu's moves and related discourse functions in the section 'Applications' where a special type of electronic manual is proposed.

Table 1: Moves and discourse functions of research articles (Nwogu 1997:125)

Move	***Discourse function***	***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	

Faber and Mairal (1999) and López Rodríguez (2002)

Faber and Mairal apply the Functional-Lexematic Model (FLM) to study the primary lexicon of English verbs. They propose a kind of lexical organization based on the well-known distinction between syntagmatic and paradigmatic relationships or the complementary principles of combination and selection (Saussure 1916; Lyons 1977:241). Syntagmatic relations are those holding between elements which co-occur in linear sequences while paradigmatic relations are those based on the potential of occurrence of elements in such combinations. According to Faber and Mairal (1999:80), the paradigmatic axis of the FLM lexicon not only codifies how terms are arranged on the axis of selection, organizing them onomasiologically in a hierarchy of domains and subdomains, but is also a determining factor in their syntactic and combinatorial possibilities. These lexical domains are defined as "The set of lexemes which together lexicalize all or part of a conceptual domain" (ibid.:59). The most prototypical verbs from a semantic point of view are those which have the largest

combinatorial potential. Faber and Mairal (ibid.) propose the following lexical domains: EXISTENCE, MOVEMENT, POSITION, CONTACT, CHANGE, PERCEPTION, COGNITION, FEELING, SPEECH, SOUND, LIGHT, POSSESSION and ACTION.

López Rodríguez (2002) bases her study of English verbs in a corpus of 156 abstracts of articles on oncology with an *IMRAD* (Introduction, Methods, Results and Discussion) structure on the lexical domains proposed by Faber and Mairal. These abstracts were selected at random from the bibliographical database MEDLINE and tagged to distinguish the four main rhetorical sections of research articles which are summarized in the abstracts: *Introduction, Methods, Results* and *Discussion*. López Rodríguez' study examines how verbs activate conceptual areas in these rhetorical sections. In the present study, the same methodology is applied to examine how verbs activate conceptual areas in the different sections of research articles.

Further aspects of the methodology applied will be explained in the next section. López Rodríguez' results will be discussed in more detail in the section 'Comparison with López Rodríguez'.

Methodology

Text selection and corpus description

In order to ensure the homogeneity of the corpus and that the articles reflect the best practice in scientific writing, the following criteria were applied in the selection of the articles:

(1) Citation index of the journal;
(2) Subject of the journal: oncology;
(3) Availability of complete articles;
(4) Introduction, Methods, Results and Discussion or IMRAD format of the article;
(5) Subject of the article: breast cancer;
(6) Date of publication.

The first and second criteria were applied using the *Journal Citation Reports* web page (JCR Web; http://jcrweb.com/jccr_selection.pl, last accessed October 15, 2006). The JCR Web assesses specialized journals according to their citations in more than 8,400 journals published by over 3,000 publishers worldwide. It provides a list of the most cited journals according to citation index and subject of the journal as defined by the JCR Web. Only three journals met the first three criteria, and these are therefore the journals used in the present study: *Carcinogenesis*; *Cancer Epidemiology, Biomarkers and Prevention*; and the *Journal of the National Cancer Institute*.

Ten articles on breast cancer were selected from each journal on the basis of the articles' keywords. Only texts on the subject of breast cancer were chosen to facilitate comprehension as the researcher in this case is not a medical expert on oncology. The resulting corpus has a total of 203,611 tokens (12,903 types), and includes:

- Ten research articles from *Carcinogenesis*, seven from the year 2000, and three published in1999;

- Ten research articles from *Cancer Epidemiology, Biomarkers and Prevention*, all published in 2000;
- Ten research articles from the *Journal of the National Cancer Institute*, three published in 2001, and seven in 2000.

All articles were published between 1999 and 2001, as more recent complete articles from the selected journals were not freely available on the Internet. No distinction was made between translations or texts originally written in English by native English or non-native authors, as this information was not available. Since all articles had been published in reputable journals, it was assumed that they complied with the minimum stylistic and linguistic requirements of the journals' editorial boards (for full references of the articles see Appendix 1).

Corpus processing

The computer processing of the corpus is largely based on the methodology developed by López Rodríguez (2002:174-176). The Brill POS Tagger was used to tag the texts of the corpus. This is a free English tagger provided by the research group for natural language processing of the Spanish National University for Distance Learning (UNED) [2]. Tags which do not refer to verb categories were eliminated by means of the Wordsmith Tools' Text Converter, which is a tool that can be used to replace strings of characters in text files.

The corpus was divided into five subcorpora according to the different sections of medical research articles: *Abstract* (10,191 t), *Introduction* (120,376 t), *Methods* (258,419 t), *Results* (192,897 t) and *Discussion* (281,335 t).

Using the *Concord* component of *Wordsmith Tools*, concordances were generated for the verbs. As the purpose of this study was to classify verbs according to their meaning in lexical domains, verbal forms of *have* and *be* were eliminated when used as auxiliary verbs. Participles which were part of a phraseological unit, such as *advanced* in, for example, *advanced breast cancer*, were eliminated as well since these participles function as adjectives rather than verbs.

Semantic classification of the verbs: lexical chain generation

For the semantic classification of the verbs extracted from the five subcorpora, I used the lexical domains identified by Faber and Mairal (1999): EXISTENCE, MOVEMENT, POSITION, CONTACT, CHANGE, PERCEPTION, COGNITION, FEELING, SPEECH, SOUND, LIGHT, POSSESSION and ACTION. They provide a hierarchical classification of English verbs structured according to the genus or conceptual label in their definitions. This hierarchically structured lexicon of verbs and their definitions (Faber and Mairal ibid.) was the basis of the classification in this corpus analysis. Some of the more general verbs (for example, "find", "observe", "show") analyzed in this corpus study can be directly found in the hierarchy proposed by Faber and Mairal, more specific verbs, such as *diagnose*, were classified by the researcher according to the definitional structures proposed by the same authors.

To identify the conceptual areas activated in each section of the research article, the verb frequencies were exported to Microsoft Excel. Verbs with related meanings, i.e. those grouped under one of the above-mentioned lexical domains, were included in the same column forming a lexical chain. In each section (*Abstract*, *Introduction*, etc.), each lexical chain amounts to a percentage which is the quotient of the total number of verbal lexemes

belonging to this lexical chain and the total number of tokens in the section (for a list of lexical domains and the verbs they include see Appendix 2) [3]. For example, in the section *Discussion*, the second most important lexical chain of verbs is PERCEPTION, and the sum of the semantically-related verbal lexemes represents 1.03% of all the lexemes, including verbs and all other categories, of this section. The total sum of the percentages of the column (1.03%) represents approximately the relative importance of the conceptual area PERCEPTION in the *Discussion* section (see Table 2).

Table 2: Verb activation of the lexical domain PERCEPTION in the Discussion *section*

PERCEPTION	
find	0.2
show	0.18
observe	0.18
detect	0.07
examine	0.06
indicate	0.05
appear	0.04
identify	0.04
note	0.04
see	0.04
confirm	0.03
characterize	0.02
diagnose	0.02
present	0.02
reflect	0.02
seem	0.02
Total	1.03

Results

Table 3 gives an overview of the conceptual organization of each section. Clear differences can be seen in the activation of conceptual areas in the different sections of the medical research articles.

Table 3: Activation of conceptual areas in the different sections of research articles

Introduction	%	Methods	%	Results	%	Discussion	%	Abstract	%
EXISTENCE	1.66	ACTION	1.31	EXISTENCE	1.71	EXISTENCE	1.9	COGNITION	1.25
COGNITION	1.14	COGNITION	0.85	PERCEPTION	1.17	PERCEPTION	1.14	EXISTENCE	1.12
ACTION	0.89	POSSESSION	0.82	COGNITION	0.71	COGNITION	0.88	ACTION	0.89
POSSESSION	0.83	EXISTENCE	0.77	CHANGE	0.7	CHANGE	0.68	PERCEPTION	0.9
PERCEPTION	0.84	PERCEPTION	0.54	POSSESSION	0.66	ACTION	0.63	CHANGE	0.79
CHANGE	0.74	POSITION	0.4	ACTION	0.63	POSSESSION	0.6	POSSESSION	0.63
SPEECH	0.45	SPEECH	0.37	SPEECH	0.18	SPEECH	0.33	SPEECH	0.17
POSITION	0.14	CHANGE	0.35	POSITION	0.18	POSITION	0.12	POSITION	0.26
MOVEMENT	0.09	MOVEMENT	0.29	MOVEMENT	0.06	MOVEMENT	0.07	MOVEMENT	0.06

In Table 3, the percentages are calculated by comparing the sum of the verbal lexemes of a lexical domain to the sum of all the words in the section. To visualize the activation of the conceptual areas more clearly, the data shown in Table 3 are plotted on a diagram for each

research article section below. In the diagrams, the percentages are calculated by comparing the sum of the verbal lexemes of one lexical domain to the total of verbal lexemes in the section. After that, the results of this study are compared with the results obtained by López Rodríguez in her analysis of medical abstracts.

Introduction section

In the *Introduction*, the lexical chain EXISTENCE amounts to a fourth of the conceptual areas (25%), and is followed by the following lexical fields ordered according to their prominence (i.e. frequency of occurrence): COGNITION, ACTION, POSSESSION, PERCEPTION and CHANGE.

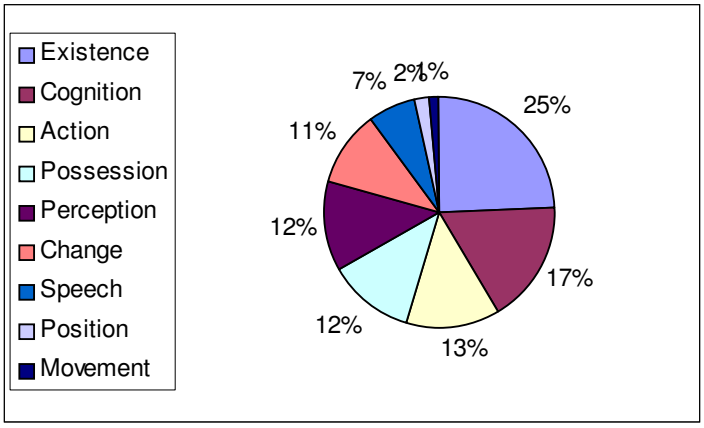


Figure 1: Activation of conceptual areas in the Introduction *section*

In the *Introduction* sections, the authors explain the hypotheses that will be tested in the study (COGNITION, see Example 1). These hypotheses are generally tested through the observation of a phenomenon (PERCEPTION, Example 2). Apart from that main function, which is usually described in a few sentences, this section also describes in quite some detail previous and related research (EXISTENCE, ACTION and POSSESSION, see Examples 3 and 4).

> Examples:
> (1) This study was undertaken [ACTION] to **investigate** [COGNITION] whether the NAT2 polymorphism is **associated** [COGNITION] with breast cancer risk…
> (2) In addition, at puberty the affected female may **show** [PERCEPTION] signs of virilization and have pubertal failure, hypergonadotrophic hypogonadism, polycystic ovaries and tall stature.
> (3) The NAT2 gene **is** [EXISTENCE] polymorphic and individuals who carry two allelic mutations **have** [POSSESSION] a slow acetylator phenotype, whereas heterozygous wild-type genotypes **have** [POSSESSION] an intermediate acetylator phenotype…
> (4) The other agent, EB 1089, has been **used** [ACTION] in the chemotherapy protocol **utilizing** [ACTION] human breast cancer cells in athymic mouse.

The verbs that are used most often (>0.1%) in the *Introduction* section are: *be, associate, investigate, know, use, have, include, find, show* and *increase* [4].

Methods section

The most representative conceptual area of the *Methods* section is ACTION (24%, see Figure 2). This clearly differentiates this section from the other sections, where EXISTENCE is the

most important lexical chain. In the *Methods* section, ACTION (see Examples 5 and 6) is followed by COGNITION (Example 6), POSSESSION (Example 6), EXISTENCE and PERCEPTION.

> Examples:
> (5) Two separate series (strata) were **used** [ACTION] to overcome the need to **increase** [CHANGE] thresholds of statistical significance when **carrying out** [ACTION] multiple tests.
> (6) For each set of treatments, we **performed** [ACTION] and **averaged** [COGNITION] three fluctuation analyses at three separate times by use of the same starting populations of cells **recovered** [POSSESSION] from freezer stocks.

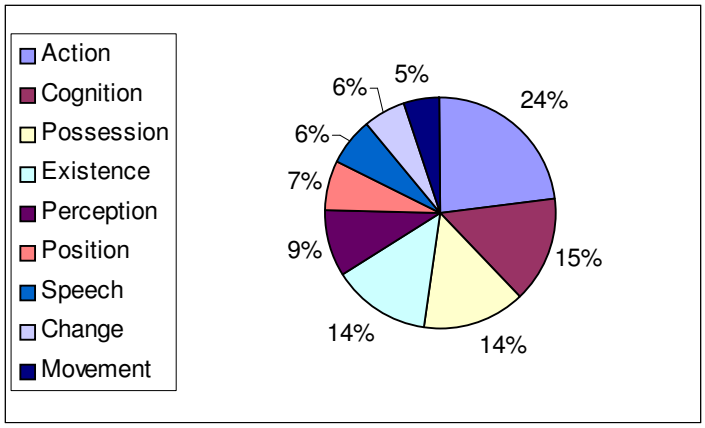


Figure 2: Activation of conceptual areas in the Methods *section*

In the *Abstract* and *Introduction* sections, the conceptual area ACTION is third in importance, and in *Results* and *Discussion*, ACTION occupies the fifth and sixth place respectively. The fact that ACTION is ranked first in *Methods*, is an indication that the *Methods* section has a specific rhetorical function which is different from the other sections of the research article, namely that of describing the data collection and analysis procedures. This specific function is also the reason why lexical chains that hardly appear in other sections become more important here, such as POSITION (see Example 7) and MOVEMENT (Example 8).

> Examples:
> (7) Transduced clones were **selected** [POSITION] first for expression of the **enhanced** [CHANGE] green fluorescent protein and then for the **reduced** [CHANGE] accumulation of mitoxantrone by flow cytometry.
> (8) Adducts were **separated** [MOVEMENT] by thin layer chromatography **using** [ACTION] the solvents **described** [SPEECH] in Singletary et al.

These results indicate that the *Methods* section has a specific role in the research article, to which the rest of the text hardly makes any reference, namely the description of the procedures applied, the experiments carried out, and the data collected and analyzed. The verbs most frequently (>0.1%) found in the *Methods* section are: *use*, *perform*, *analyze, assess, contain, have, include, obtain, be* and *determine*.

Results section

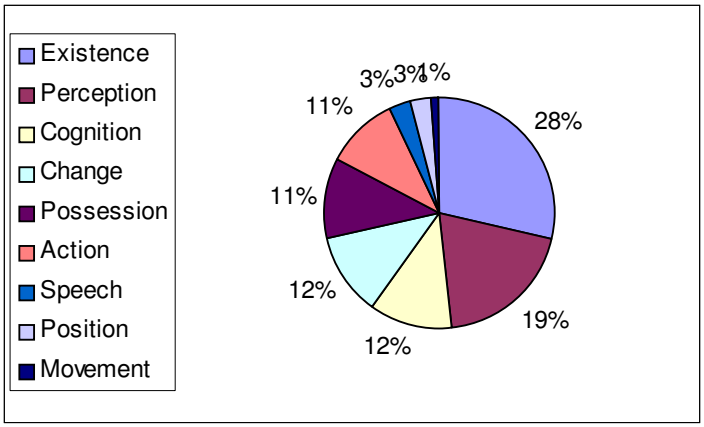


Figure 3: Activation of conceptual areas in the Results *section*

The most representative lexical chain of the *Results* section is EXISTENCE (28%), followed by PERCEPTION, COGNITION, CHANGE, POSSESSION and ACTION. The *Results* section is used to describe the data obtained during research (EXISTENCE and PERCEPTION, see Examples 9 and 10), and assert if they support the initial hypotheses (COGNITION, Example 11). The lexical chain POSSESSION (see Example 12) is activated to describe the properties of the observed phenomena.

> Examples:
> (9) There **are** [EXISTENCE] no significant differences between cases and controls in the intake of calories, fat or fruits and vegetables, but there **is** [EXISTENCE] a suggestion of higher protein intake…
> (10) No mutations were **detected** [PERCEPTION] in the control experiment.
> (11) … we **compared** [COGNITION] allele frequencies between cases and controls to **look** [PERCEPTION] for any association with breast cancer risk.
> (12) This result **indicates** [PERCEPTION] that inhibition of both Bcrp1 and P-gp by GF120918 **has** [POSSESSION] a strong effect on uptake of topotecan administered orally…

The verbs most frequently used in the *Results* section are (>0.1%): *be, detect, observe, show, associate, compare, increase, affect, have* and *use*.

Discussion section

In the *Discussion* section, the same conceptual areas as in the *Results* section prevail. This is probably due to the fact that the results presented in the previous section are interpreted in the *Discussion* section: EXISTENCE (see Examples 13 and 15), PERCEPTION (Example 14), COGNITION (Example 15), POSSESSION (Example 13), CHANGE and ACTION.

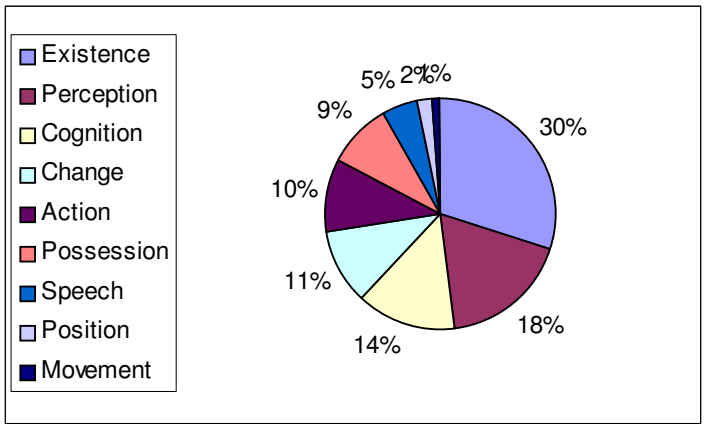


Figure 4: Activation of conceptual areas in the Discussion

Examples:
(13) Our results **suggest** [SPEECH] that among the phenols **contained** [POSSESSION] in brown rice, tricin might **be** [EXISTENCE] a prime candidate nutriceutical with colon or particularly breast cancer chemopreventive activity.
(14) Modest inverse associations between alcohol consumption and hormone levels were **observed** [PERCEPTION] with testosterone and SHBG.
(15) Susceptibility to mutagen treatment MNNG (an alkylating mutagen) is **known** [COGNITION] to **induce** [EXISTENCE] single-strand DNA breaks as **detected** [PERCEPTION] by the Comet assay (20,21).

Not surprisingly, the verbs that prevail in this section are very similar to the ones found in the *Results* section (>0.1%): *be, induce, find, observe, show, associate, compare, increase* and *use*.

Abstract section

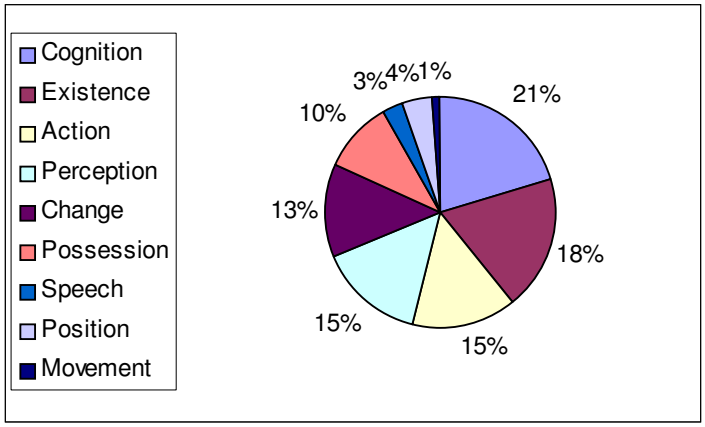


Figure 5: Activation of conceptual areas in the Abstract

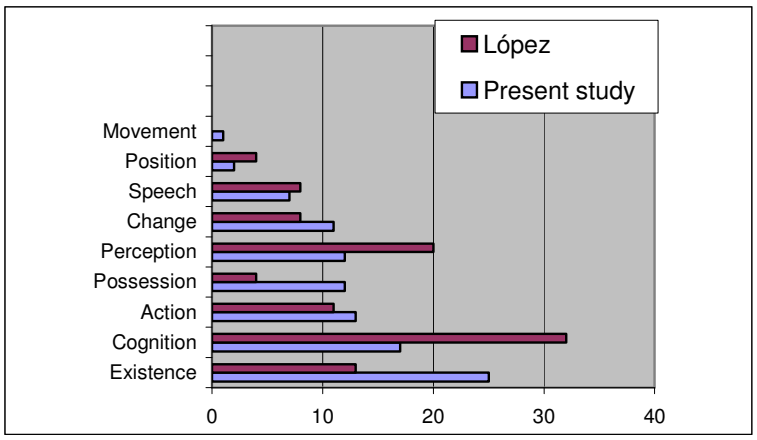
If we consider that the abstract reflects all the other sections of the research article, all the prevailing lexical chains of these sections should be present in the abstract. The most important conceptual area of this section is COGNITION, followed by EXISTENCE, ACTION, PERCEPTION, CHANGE and POSSESION. The percentages of the four largest lexical domains are very similar (21%, 18%, 15% and 15%). Moreover, in every section of the body of the

research article at least three of these four lexical domains are among the four most prominent ones.

The results of this corpus analysis show that each section of the research article is of similar importance in the abstract. This is in line with Swales' (1990:179) suggestion that abstracts are 'distilled' versions of the whole text. This means that only the essential and most interesting information is included in the abstract. Its distilled quality is shown by the relatively low percentages of the lexical chain POSSESSION, which is used to describe properties, and CHANGE, which is often applied in detailed descriptions. This is very important if we take into account that readers are likely to start by reading the abstract and many may not read the whole article.

Comparison with López Rodríguez

As mentioned earlier, the methodology applied in the present study was elaborated by López Rodríguez (2002) to analyze verbs in abstracts of medical research articles. She subdivided each abstract according to what was explained of each section (*Introduction*, *Methods*, *Results* and *Discussion*) in each abstract. In the current study I have divided complete articles into their respective sections for analysis. In the previous section it has become clear that a special relationship exists between the complete articles of the corpus of texts and their abstracts. Given that an abstract is a condensed summary of the whole body of an article, it seems reasonable to assume that López Rodríguez's results could be extrapolated to the whole article. For this reason, a comparison has been made between her results and the ones obtained in this study. López Rodríguez's corpus consists of 156 abstracts of articles on oncology selected at random from the bibliographical database MEDLINE (43690 t). The methodology she applies is the same as the one applied in the present study. This means that the percentages of verbal lexical chains are calculated by comparing the sum of the verbal lexemes of a lexical domain to the sum of all the words in the each rhetorical section of the abstracts, in the same way as in the present study the percentages are compared to all the words in each section of the article. Therefore, the percentages of the two studies can be compared. It should be taken into account, however, that abstracts and articles also tend to differ in terms of style. The need for conciseness in the abstracts may lead, for instance, to a higher frequency of content words, and therefore the percentage of verb forms could be expected to be higher in abstracts than in articles. Nevertheless, the relative prominence of the different lexical chains in each section of the abstract/article could be expected to remain the same. In any case, this is only an exploratory study intended to shed more light on the relationship between abstracts and complete articles, and further research must be carried out to support the results of this study.



Bar chart 1: Comparison of conceptual activation in the Introduction *section*

Regarding the *Introduction* section, there are clear differences between the data of López Rodríguez and the data obtained in this study (see Bar chart 1). In the first place, EXISTENCE (25%) is the most representative lexical chain in the results of this analysis, but in López Rodríguez it is COGNITION (32%). This is probably due to the fact that an abstract is a summarized and condensed text in which the rhetorical section that corresponds to the *Introduction* can only define the hypothesis (COGNITION). The research context (EXISTENCE) is not described due to space constraints. However, in complete articles, the *Introduction* section tends to describe previous research and studies carried out by other authors and the hypothesis only extends to a few lines.

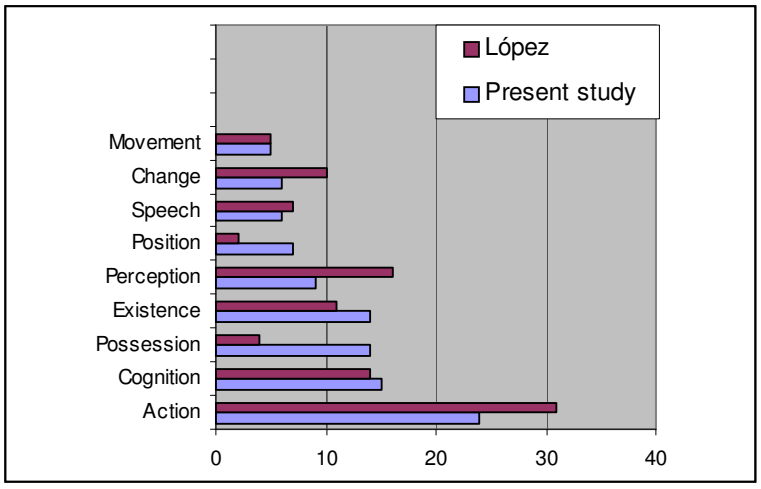
The same argument explains the different percentages of the lexical chains PERCEPTION (20% in López and 12% in the present study), POSSESSION (4% and 12% respectively), ACTION (11% and 13%), and EXISTENCE (13% and 25%). In a complete article there is more space available to describe the research context, the way in which previous studies have been carried out (EXISTENCE and ACTION). Verbs of the lexical chain POSSESSION are used to go into even more detail (see examples 16, 17 and 18). The verb *provide*, for example, is often used to explain what is important about a previous study and which parts are applicable to the research at hand (see example 17).

Examples:

(16) However, these techniques use (ACTION) pooled cell populations and, thus, are unable to provide (POSSESSION) critical information about intercellular differences in DNA damage and repair.

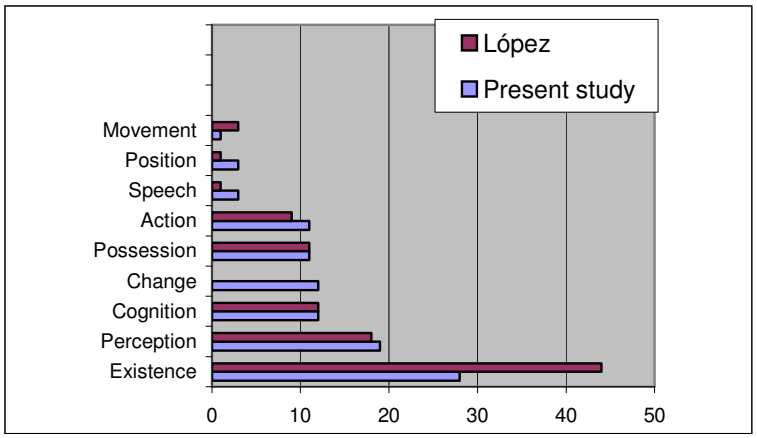
(17) This cohort has provided (POSSESSION) a wealth of data about active smoking (29,32,33).

(18) Candidates for such low-risk genes include (POSSESSION) those involved (POSSESSION) in cancer predisposition pathways….

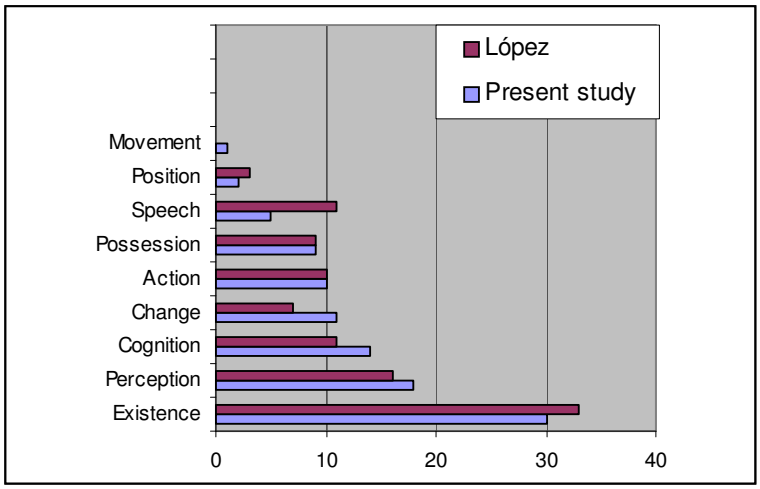


Bar chart 2: Comparison of conceptual activation in the Methods *section*

Bar chart 2 shows that ACTION is the most important lexical chain in both López Rodríguez (31%) and the data presented here (24%). However, there are quite a few differences in the lexical chains POSSESSION (present study: 14%; López: 4%), PERCEPTION (9% and 16%) and CHANGE (6% and 10%). Just like in the *Introduction*, the higher prominence of the conceptual area POSSESSION in the *Methods* section in the data of this investigation is due to the fact that there is more space available to describe all the steps of the experiments carried out in the study. Because the conceptual area POSSESSION increases, the relative importance of the conceptual areas PERCEPTION and CHANGE decrease in the data obtained in the present study as compared to López Rodríguez.



Bar chart 3: Comparison of conceptual activation in the Results *section*



Bar chart 4: *Comparison of conceptual activation in the* Discussion

The results of both López Rodríguez and the data of this study are very similar in the *Results* and *Discussion* sections (see Bar charts 3, 4). For example, the differences in the fields of CHANGE and EXISTENCE comparing López Rodríguez and the present study are due to the same reason of space limitations in the former. The higher prominence of speech in the *Discussion* both in López and the present study, although more obvious in the former, can be explained by the fact that in this section the results are explained and discussed upon.

What is especially striking when we compare the findings of the present study with those obtained in López Rodríguez (2002) is that the differences among the percentages of the conceptual areas are greater in the latter. In other words, the results of López Rodríguez seem to be more extreme than those of the present investigation and the rhetorical function of each section seems to be clearer. This seems to be justified by the fact that an abstract is not only a perfect summary of the body of the article but, as suggested by Swales (1990), it is also a 'distilled' version of it: only the most important and interesting information is included. In the body of an article all topics can be discussed in detail and hence explanations and reference to other sections can be included. Consequently, although there are relative differences in the presence of conceptual areas according to article section, all conceptual areas can be found in all sections, and the predominance of one or other is not that pronounced in a complete research article.

Other possible comparisons could be between the rhetorical sections of the abstracts and the other sections of the articles in the current study, or between the abstracts of the current study and López Rodríguez's data. To carry out these comparisons, the abstracts of the present study would have to be tagged in the same way as López Rodríguez's. For the time being, the abstract has been considered as one of the sections of the complete article, but further research will look into these possibilities.

Applications

In general as well as in specialized language the choice of adequate words is crucial. Writing or translating a research article on a specialized topic such as medicine is not only difficult because of its specialized content and terminology, but also because of the need to apply specific writing conventions which may change from one academic community to another.

The results of this study constitute a first attempt to analyze a specific word group (verbs) in each of the sections of the medical research article. The connection between the lexical field, the research article section and the specific use of a verb can help the writer or translator in his/her choice of vocabulary. However, to make the results of this analysis of use to writers and translators, the information would have to be organized different and made accessible through other means. One way of doing this would be to produce an electronic manual that assists medical professionals and translators in selecting the most common and appropriate verbs for each section of the medical research article [5]. This manual could be organized by article section (*Abstract*, *Introduction*, *Methods*, *Results* and *Discussion*) in the first place and then by lexical domain. This manual could also be accessible through an alphabetical list of verbs, but this list should always lead the user to an entry that relates the verb to its lexical domain, the relevant section of the research article, and to other relevant verbs.

To relate the lexical characteristics of the medical language, in this case main verbs, to the rhetorical functions of the different sections of the research article, the entry might also contain information on the discourse functions described in Nwogu (1997). In this way, the manual would be enriched with information on macrostructure. For the user who is not used to linguistic terminology, and for illustration purposes, the entry could have examples of usage extracted from the corpus. A sample record is provided in Figure 6.

Introduction

Discourse functions

1. To present background information
    a. Reference to established knowledge in the field

    The development of effective radiotherapy and chemotherapy regimens for the treatment of HD 3 has resulted [CHANGE] in large numbers of long-term survivors. (cebp1b)

    Since vitamin D exhibits [PERCEPTION] marked cell-differentiating activity, it can be used [ACTION] as a possible chemopreventive agent (4,5). (jnci5b)

    b. Reference to main research problems

    The relationship between NAT2 polymorphisms and breast cancer carcinogenesis is not clearly understood [COGNITION]. (cage4b)
2. To review related research
    a. Reference to previous research

    Epidemiological studies have identified [PERCEPTION] several factors affecting [CHANGE] lifetime exposure that increase [CHANGE] the risk of breast cancer. (cage6b)

    c. Reference to limitations of previous research

    However, there are [EXISTENCE] few data on the molecular changes that accompany [POSSESSION] lung (12) or breast cancer after treatment for HD…. (cebp1b)
3. To present new research
    a. Reference to research purpose

    …this type of data might provide [POSSESSION] further insight into the contribution of various etiological factors and carcinogenic pathways compared [COGNITION] with de novo tumors. (cebp1b)

    b. Reference to main research procedure

    To identify [PERCEPTION] a potential mechanism, we assessed [COGNITION] the relationships among ADH3 genotype, alcohol consumption, and plasma levels of steroid hormones in a subset of these women. (cebp5b)

Lexical domains

1. EXISTENCE
2. COGNITION

Verbs: *assess, associate, compare, demonstrate, determine, estimate, investigate, know, relate, study*

Study

Definition
To think carefully about something, looking at it in your mind from different perspectives in order to understand it better, observing and analyzing it.

Concordances

enzyme-altered focal lesions have been studied for their relevance to carcinoma development
as preventive therapies are being studied and becoming available for that population (1
Rebbeck et al. (19) have studied this polymorphism in carriers of BRCA1 mutat
e do not know whether the populations studied were similar with respect to these two
transgenic rats have been developed to study carcinogenesis. Rats containing an albumin
these considerations, we decided to study the potential colon- and breast tumor-suppress
transferase gene has been employed to study regulation of GST-P transcripts in rat liver
umber of serum repositories available to study the effects of pregnancy serum components

Figure 6: Sample record of the Introduction *section*

In Figure 6, the entry consists of several types of information. Firstly, we find the discourse functions of the section (as proposed by Nwogu 1997), for example "to present background information". These discourse functions are then illustrated by examples taken from the corpus which will help the users to understand how each discourse function can be realized in a research article.

Secondly, the entry presents the lexical domains in order of importance, in this case (*Introduction* section) starting with the domain EXISTENCE. Within each lexical domain we find the verbs that belong to it. The verbs (for example, *study* in the lexical domain COGNITION), could be hyper-linked to their definition and a list of concordances. These concordances show how the verb functions in the context of the article section and its collocates.

Conclusion

The corpus analysis presented here focused on the lexical aspects of scientific language, more precisely, on the verbs used in research articles with an IMRAD organization. In this study, the method applied integrates corpus linguistics, the research in lexical semantics carried out by Faber and Mairal (1999) and Nwogu's (1997) discourse analysis. The semi-automatic method developed by López Rodríguez (2002) to study verbs in the rhetorical sections of medical abstracts has been applied to the different sections of complete research articles. The results indicate that in each section different lexical domains prevail that reinforce the rhetorical functions of each section. The comparison of the results of López Rodríguez with the results of the present study sheds light on the relationship between an abstract and the body of an article and confirms Swales's (1990) statement that the abstract is not only a summary but also a 'distilled' or 'purified' version of the whole text. The results of studies like this can be used in a practical way. For instance, the type of manual proposed above may

serve as a valuable tool to assist medical experts and translators in writing and translating research articles.

Author's address
Callejón de Santo Domingo 6, 3B
18009 Granada
Spain
Phone: +34-958-224360
arianne(a)ugr.es

Notes

1 This research is part of the research project BFF2003-04720, funded by the Spanish Ministry of Education.
2 The Brill POS Tagger can be found at http://rayuela.ieec.uned.es/cgi-bin/ircourse/bril.perl
3 The quotient of the total number of lexemes and the total number of words of the selected corpus is a number provided automatically by *Wordsmith Tools*.
4 Only verbs that represent more than 0.1% are included.
5 An example of an electronic dictionary based on the results of this study can be found in *Redactar y traducir artículos de investigación: un programa de software* (Reimerink 2006) where a software program for writing and translating medical research articles is included.

References

Faber, Pamela and Catalina Jiménez (2002) *Investigar en terminología*, Granada: Comares.
Faber, Pamela and Ricardo Mairal Usón (1999) *Constructing a Lexicon of English Verbs*, Berlin/New York: Mouton de Gruyter.
Hoof, Henri, van (1998) 'The language of medicine: a comparative ministudy of English and French', in Henry Fischbach (ed) *Translation and Medicine*, American Translators Association, Scholarly Monograph Series, Vol. X, Amsterdam/Philadelphia: John Benjamins, 49-65.
López Rodríguez, Clara Inés (2000) *Tipología Textual y Cohesión en la Traducción Biomédica Inglés-Español: un Estudio de Corpus*, Unpublished MA Dissertation, University of Granada.
López Rodríguez, Clara Inés (2002) 'Extracción de información conceptual, textual y retórica en terminología: la distribución de verbos en los resúmenes de artículos experimentales', in Pamela Faber and Catalina Jiménez (eds) *Investigar en Terminología*, Granada: Comares. 167-195.
Lyons, John (1977) *Semantics*, vol. 1 and 2, London: Cambridge University Press.
Navarro, Fernando A. (2001) 'El inglés, idioma internacional de la medicina: causas y consecuencias de un fenómeno actual', *Médico Interamericano*, 20: 16-24.
Nwogu, K. N. (1997) 'The medical research paper: structure and functions', *English for Specific Purposes* 16(2): 119-38.
Pilegaard, Morten (1997) 'Translation of medical research articles', in Anna Trosborg (ed) *Text Typology and Translation*, Amsterdam/Philadelphia: John Benjamins, 159-184.
Reimerink, Arianne (2006) *Redactar y traducir artículos de investigación: un programa de software*, Unpublished MA Dissertation, University of Granada.
Saussure, Ferdinand, de (1916) *Cours de linguistique générale*, Paris: Payot.
Swales, John. M. (1990) *Genre Analysis*, Cambridge: Cambridge University Press.
Tercedor Sánchez, María Isabel (2002) 'Descripción y representación de la variación terminológica: el caso de la dimensión tipos de cáncer', in Pamela Faber and Catalina Jiménez (eds) *Investigar en terminología*, Granada: Comares. 199-214.
Tercedor Sánchez, María Isabel and Beatriz Méndez Cendón (2000) 'Fraseología y variación terminológica: estudio descriptivo en corpora biomédicos', *Terminologie et Traduction* 2: 82-100.

The web pages of the medical journals used in the present study are (last accessed 23rd January, 2007):
Carcinogenesis: http://carcin.oxfordjournals.org/

Appendix 1: References of corpus articles

Carcinogenesis

Asamoto, Makoto et al. (2000) 'Transgenic rats carrying human c-Ha-*ras* proto-oncogenes are highly susceptible to *N*-methyl-*N*-nitrosourea mammay carcinogenesis', *Carcinogenesis* 21(2): 243-249.

Bianco, Tina et al. (2000) 'Tumour-specific distribution of *BRCA1* promoter region methylation supports a pathogenetic role in breast and ovarian cancer', *Carcinogenesis* 21(2): 147-151.

Clay, Carl E et al. (1999) 'Influence of J series prostaglandins on apoptosis and tumorigenesis of breast cancer cells', *Carcinogenesis* 20(10): 1905-1911.

Delfino, Ralph J. et al. (2000) 'Breast cancer, heterocyclic aromatic amines from meat and *N*-acetyltransferase 2 genotype', *Carcinogenesis* 21(4): 607-615.

Dunning, Alison M. et al. (1999) 'No association between androgen or vitamin D receptor gene polymorphisms and risk of breast cancer', *Carcinogenesis* 20(11): 2131-2135.

Healey, Catherine S. et al. (2000) 'Polymorphisms in the human aromatase cytochrome P450 gene (*CYP19*) and breast cancer risk', *Carcinogenesis* 21(2): 189-193.

Manjanatha, M. G. et al. 'DNA adduct formation and molecular analysis of *in vivo lacI* mutations in the mammary tissue of Big Blue® rats treated with 7,12-dimethylbenz[α]anthracene', *Carcinogenesis* 21(2): 265-273.

Rajeswari, N. et al. (2000) 'Risk assessment in first degree female relatives of breast cancer patients using the alkaline Comet assay', *Carcinogenesis* 21(4): 557-561.

Tan, Xingzhi, Arthur P. Grollman and Shinya Shibutani (1999) 'Comparison of the mutagenic properties of 8-oxo-7,8-dihydro-2'-deoxyguanosine DNA lesions in mammalian cells', *Carcinogenesis* 20(12): 2287-2292.

Tsutsui, Takeki et al. (2000) 'Induction of mammalian cell transformation and genotoxicity by 2-methoxyestradiol, an endogenous metabolite of estrogen', *Carcinogenesis* 21(4): 735-740.

Cancer Epidemiology, Biomarkers and Prevention

Behrens, Carmen et al. (2000) 'Molecular changes in second primary lung and breast cancers after therapy for Hodgkin's disease', *Cancer Epidemiology, Biomarkers and Prevention* 9: 1027-1035.

Caan, Bette J. et al. (2000) 'Low-energy reporting in women at risk for breast cancer recurrence', *Cancer Epidemiology, Biomarkers and Prevention* 9: 1091-1097.

Cui, Jisheng and John L. Hopper (2000) 'Why are the majority of hereditary cases of early-onset breast cancer sporadic? A simulation strudy', *Cancer Epidemiology, Biomarkers and Prevention* 9: 805-812.

Deitz, Anne C. et al. (2000) '*N*-acetyltransferase-2 genetic polymorphism, well-done meat intake, and breast cancer risk among postmenopausal women', *Cancer Epidemiology, Biomarkers and Prevention* 9: 905-910.

Hines, Lisa M. et al. (2000) 'A prospective study of the effect of alcohol consumption and *ADH3* genotype on plasma steroid hormone levels and breast cancer risk', *Cancer Epidemiology, Biomarkers and Prevention* 9: 1099-1105.

Hudson, E. et al. (2000) 'Characterization of potentially chemopreventive phenols in extracts of brown rice that inhibit the growth of human breast and colon cancer cells', *Cancer Epidemiology, Biomarkers and Prevention* 9: 1163-1170.

Mitrunen, Katja et al. (2000) 'Steroid metabolism gene *CYP17* polymorphism and the development of breast cancer', *Cancer Epidemiology, Biomarkers and Prevention* 9: 1343-1348.

Richardson, Barbara E., Jennifer D. Peck and Jennifer K. Wormuth (2000) 'Mean arterial pressure, pregnancy-induced hypertension, and preeclampsia: evaluation as independent risk factors and as surrogates for high maternal serum α-fetoprotein in estimating breast cancer risk', *Cancer Epidemiology, Biomarkers and Prevention* 9: 1349-1355.

Stellman, Steven D. et al. (2000) 'Breast cancer risk in relation to adipose concentrations of organochlorine pesticides and polychlorated biphenyls in Long Island, New York', *Cancer Epidemiology, Biomarkers and Prevention* 9: 1241-1249.

Ward, Elizabeth M. et al. (2000) 'Serum organochlorine levels and breast cancer: a nested case-control study of Norwegian women', *Cancer Epidemiology, Biomarkers and Prevention* 9: 1357-1367.

Journal of the National Cancer Institute

Freedman, Matthew et al. (2001) 'Digitized mammography: A clinical trial of postmenopausal women randomly assigned to receive raloxifene, estrogen, or placebo', *Journal of the National Cancer Institute* 93(1): 51-56.

Galper, Sharon R. et al. (2000) 'Patient preferences for axillary dissection in the management of early-stage breast cancer', *Journal of the National Cancer Institute* 92(20): 1681-1687.

Herbert, Brittney-Shea et al. (2001) 'Effects of chemopreventive and antitelomerase agents on the spontaneous immortalization of breast epithelial cells', *Journal of the National Cancer Institute* 93(1): 39-45.

Jonker, Johan W. et al. (2000) 'Role of breast cancer resistance protein in the bioavailability and fetal penetration of topotecan', *Journal of the National Cancer Institute* 92(20): 1651-1656.

Mehta Rajendra et al. (2000) 'Prevention of *N*-methyl-*N*-nitrosourea-induced mammary carcinogenesis in rats by 1(alpha)-hydroxyvitamin D5', *Journal of the National Cancer Institute* 92(22): 1836-1840.

Paik, Soonmyung et al. (2000) 'HER2 and choice of adjuvant chemotherapy for invasive breast cancer: national surgical adjuvant breast and bowel project protocol B-15', *Journal of the National Cancer Institute* 92(22): 1991-1998.

Reis, Steven E. et al. (2001) 'Cardiovascular effects of tamoxifen in women with and without heart disease: breast cancer prevention trial', *Journal of the National Cancer Institute* 93(1): 16-21.

Schiff, Rachel et al. (2000) 'Oxidative stress and AP-1 activity in tamoxifen-resistant breast tumors in vivo', *Journal of the National Cancer Institute* 92(23): 1926-1934.

Spurdle, Amanda B. et al. (2000) 'CYP17 promoter polymorphism and breast cancer in Australian women under age forty years', *Journal of the National Cancer Institute* 92(20): 1674-1681.

Wartenberg, Daniel et al. (2000) 'Passive smoking exposure and female breast cancer mortality', *Journal of the National Cancer Institute* 92(20): 1666-1673.

Appendix 2: Lexical domains and their verbs

EXISTENCE	Be, form, generate, induce, initiate, occur, result
COGNITION	Assess, associate, assume, compare, consider, demonstrate, determine, evaluate, hypothesize, investigate, know, relate, study
ACTION	Act, carry out, cause, conduct, control, perform, test, treat, use
POSSESSION	Contain, give, have, include, involve, obtain, provide, receive
PERCEPTION	Appear, confirm, characterize, diagnose, detect, examine, find, identify, indicate, note, observe, present, reflect, reveal, see, seem, show
CHANGE	Affect, alter, change, decrease, enhance, increase, influence, modify, reduce, vary
SPEECH	Conclude, describe, explain, propose, report
POSITION	Assign, base, exclude, position
MOVEMENT	Direct, enroll, follow