Infopragmatics: an efficient method for information retrieval

Ibarra Rafael and Ballesteros Silvia

National Autonomous University of Mexico.

Abstract: Based on a linguistic algorithm, supported by an uncertainty theorem, the *infopragmatics* is a new method that offers an efficient solution, but not limited, to those Spanish speaking users who try to get the most useful information from academic databases which contents is in English. Presents a search analysis, an application of the language of levels understanding table, brief considerations on the ambiguity of the term "relevance" and statistical reasons to put *infopragmatics* into action at our National University Library System.

……………………………………………………………………………………

Keywords: infopragmatics – information retrieval (IR) – algorithm – uncertainty theorem – pragmatics – linguistics – Spanish speakers.

Palabras clave: infopragmática – recupearción de información (RI) – algoritmo – teorema de incertidumbre – pragmática – lingüística – nativo hablantes del español.

……………………………………………………………………………………

- **Purpose** Offers an efficient method for information retrieval from commercial academic databases.
- **Design/methodology/approach** Consists of three action points: a linguistic algorithm, a linguistic storm and an uncertainty theorem.
- **Findings** Evidence of chaotic interfaces, dim help and semantic cracks in ISs
- **Research limitations/implications** Analysis made on Spanish speakers users.
- **Practical implications** Applicable to any language if the target language is English.
- **Originality/value** No previous similar linguistic three step method approach for Native Spanish Speakers

Introduction

Information Retrieval is an unfinished challenge faced by today's librarians in order to give satisfaction to their users with pertinent information. There are several reasons that make this problem unsolvable, among others: information systems chaotic interfaces; users multiple categories and the unavoidable mismatch between controlled and natural vocabulary. These three critical aspects and their underlying grounds lead users in general to experience an uncertainty phase that overwhelms and inhibits them to get the information they need from information systems (ISs), also known as data bases (DB).

Furthermore, though there have been several alternatives given by ISs providers and uncountable researchers' theoretical and practical contributions and robotic solutions, users' demands for a suitable method to their information satisfaction have not been decreased sufficiently: they are still looking for an appropriate method. *Infopragmatics* is offered to relieve those crucial aspects by means of a series of linguistic tools: a *linguistic algorithm*, an *uncertainty theorem* and a *linguistic storm* (*lingstorm*) – that result in an effective approach to get an efficient and human interactive alternative to satisfy users' information demands, specially for, but not limited to, those Non Native English Speakers (NNES).

Before going into deep, it should be mentioned that though there is a variety of DB that provide data in several formats: text, image, sound, video, and their possible combinations, the present paper will only consider commercial ISs companies with academic information, such as the ones offered by ProQuest, EBSCO and Elsevier because they are the ones used by the involved academic community. Google, Yahoo and the like are not considered in this study due to their nature as search engines, commercial sponsored interests and aims, though they are information alternatives many user take.

2. Information systems chaotic interfaces

The commercial ISs above mentioned present chaotic interfaces search pages and a quick access cumbersome help key (F1) that clearly show how unsuitable they might be for the numerous users' kinds, but especially for a novice user whose language is not English. For example, the DB **Academic Search Premier**, from Ebsco, interface presents 35 elements, that become active at the ease of the user with a click, in the initial page, and many of them are of not very attractive for a novice user. Some of such elements are: *References available, scholarly journals, smart text searching, apply related words, visual search, image quick view, cover story(7). On its side,* ***Science Direct****, from Elsevier, presents, more than 30 elements, among which there are: please sign up (help us to improve), logged in (favourite books/journals), alerts, submit an article, read more (science direct enhancements). ProQuest, presents 16 elements, and the ones that may be of no use for the novice user are related to the privacy policy, terms and conditions, contact. It is true that most of the elements in the three ISP are of great help for expert users, whose native language is English.*

According to the last reported inform, the top three consulted databases at the UNAM in 2006 were *Academic Search Premier* (25,723 consultations), *Elsevier Science* (7753 consultations), *PsychINFO* (5472 consultations). After reviewing the interfaces of each of these providers, it was found that they fit adequately for a certified librarian or user by the companies, as well as for a robot to read and select the check lists presented in an instant, but they are definitively not easy to deal with for a standard librarian or user whose language is Spanish or some other language different than English.

The lack of uniformity and troublesome interfaces of the ISs prevent several users to use them and prefer popular search engines that offer a suitable interface of no more than nine elements in English and five in Spanish language.

2. 1 Kinds of users and their language

Users are unique in their ideas, thoughts, and needs, so the way in which they express their information queries is as different as can be. On the other side, Spanish speaker users, either students or librarians, are not one of a kind and, internationally, share characteristics that can be summed up in the following, but partial categories referred by Stubinz and Whighly (2003): *Pip*, the impatient; *Odysseus*, the dogged; *Ishmael*, the exploratory; *Hamlet*, the confused; *Ophelia*, the deranged and, concerning the Spanish language, *Don Quixote*, the idealist, can be added as well. Though this is a limited variety of users, it can be generally appreciated that, besides their human emotional features, users speak different languages and are potential to experience frustration no matter their language: English, Greek, Hebrew, Danish and Spanish.

With this scope, it is not difficult to imagine the effects that the merge between these kinds of users, different languages and commercial ISs' chaotic interfaces may cause. Among the possible effects derived from the lack of method for novice users there are two capital ones: frustration and uncertainty.

In current times it is evident that researchers at all levels may have at hand a great number of strategies with numerous modifications and concerns to know, manage and improve the manners and channels of user services. However, as Katsirikou and Skiadas (2001) mention, there are 23 processing actions that comprise the opening and the closing dialog in an information request that go from finding the appropriate electronic resource to indirectly and unwittingly provide personal information on one's activities, no matter the language. Expert users may as well avoid those elements that are of no interest for them, but what can novice users do?

2.2 Spanish Speakers Experience with ISs in English language

The National Autonomous University of Mexico Library System (UNAM LS) offers access to more than 198 databases to its users concerning their fields of study and 91% of these databases are in English language. Its services are given in more than 140 libraries, but the analysis presented and referred in this work took place at the Central Library (CB), where the interaction between a librarian and a user is both: spoken and in Spanish. Nevertheless, the interaction between humans and the ISs must be written and in English, which already sets a series of probable problems: bad spelling, wrong affixation and regular use of natural language.

Frequently, novice users verbally require their information needs in Spanish and the librarian asks them to formulate their request by using the terms in written English. At this point, it is unavoidable to step in the field of interaction, "the process by which samples of the target language become available to the learner for interlanguage construction..." (Ellis, 1990) and so, both users and librarians act in accordance to the circuit of asking and offering pertinent information respectively. Nevertheless, failures are common because "lack of shared background on the part of the interlocutors interact[ed] with their lack of shared linguistic code". (Varonis, E. and Gass, S.1985).

Every year more than 9,000 graduate students and more than 37,000 undergraduate enrol as students at the National University of Mexico[1]. Despite the fact that many librarians hold their posts during several years and offer annual courses related to databases use, thousands of users are mostly new to ISs.

[1] http://www.estadistica.unam.mx/agenda/agendas/2008/disco/xls/124.xls

In the Spanish speakers experience, at the moment, novice users can not easily cop with technical difficulties: ISs chaotic interfaces and the English controlled vocabulary. Thus, the interaction between librarian and user rises as an essential linguistic scenario where several phenomena take place, either to end well or not, since “La véritable substance de la langue n'est pas constituée par un système abstrait de formes linguistiques ni par l'énonciation-monologue isolée, ni par l'acte psycho-physiologique de sa production, mais par le phénomène social de l'interaction verbale, réalisée à travers l'énonciation et les énonciations. L'interaction verbale constitue ainsi la réalité fondamentale de la langue”. (Bakhtine 1977:136).

3. Controlled vocabulary and Natural language

One of the oldest antecedents of controlled vocabulary and natural language is the dispute between two groups: the *anomalists* and the *analogists*. Generally, the first group stated that human language should be spoken and lived in the way in which it is articulated; and the second one held that humans should express their ideas under a common norm. As the years passed, the groups turned to be prescriptive and descriptive. (*Dictionary of the History of Ideas* (2003)).

This dispute explains, in some way, a semantic gap that users experience when they are using their natural language in order to find controlled vocabulary information for their formal research. They are not aware that the semantic cracks that exist between ISs and them (Ibarra, 2007) is, in most cases, the reason why after a fruitless information search they end feeling uncertain and frustrated. They would certainly prefer feeling “found” than feeling “lost” if they knew how.

Users want to satisfy their information needs in the best possible way; they need to pass from the natural language side to the controlled vocabulary one, by means of a linguistic bridge, but do not know any and do not know how.

Two basic linguistic tools, either printed or electronic, that represent the needed bridge are subject headings books, specialised dictionaries and thesauri. However, not all the DB offer this kind of help. For example, the Medline DB offered by **ProQuest** warns the users: *There are no thesauri available for your selected databases. For a list of controlled terms go to the Indexes tab and choose the Descriptors Index*; the Biological Sciences DB, offered by the same provider, allows Life Sciences Thesaurus in English; Life Sciences Thesaurus in Spanish Beta version; Life Sciences Thesaurus in French Beta version; and Taxonomic Terms. By its part, Health Source: Nursing/Academic Edition DB by **EBSCO**, grants the Merriam-Webster's Medical Desk Dictionary, Indexes, but no thesaurus nor MeSH; the Academic Search Complete DB, by EBSCO, displays a list of Subject terms. And, finally, Science Direct DB, by **Elsevier**, does not provide any dictionary, MeSH, or thesaurus. Where can novice users find some help to pass from natural language to controlled vocabulary?

3.1 Relevance

One striking and screened fact that in some way contributes to users’ uncertainty is the adjective *relevant*. On the one hand, the IR results presented by the DB are offered, in several cases, based on the *relevance* they represent according to the number of times that a term appears on a document or, on a prefixed chronological order, which can be from the oldest document to the most recent or vice versa. On the other hand, and attending the Oxford Advanced Genie, the term *relevant* means, *closely connected with the subject you are discussing or the situation you are thinking about.*

The relationship between these two divergent assumptions may lead a restless reader to consider a third one like that offered by Dreyfus (1992) “Our everyday coping skills and the global familiarity they produce determine what counts as the facts and the relevance of all facts and so are already presuppose in the organization of the frames and slots Good Old Fashion Artificial Intelligence uses for representing these facts. That is why human beings cop more easily and expertly as they as they learn to discriminate more aspects of a situation, whereas, for data bases of frames and rules, retrieving what is relevant becomes more and more difficult the more they are told". On these bases, it would be useful to know how to discriminate those data that blur the pertinent ones and a way.

Among the numerous efforts that have been done to retrieve pertinent data Ide and Véronis (1998) remark AI-base methods, Symbolic methods, connectionist methods, knowledge based-methods, machine readable dictionaries, computational lexicons, corpus based methods (automatic sense-tagging and overcoming data sparseness. However, the term *relevance* used by the ISs providers to present the results of their Information Retrieval Systems unwillingly masks the meaning *repetition* and *anaphora,* so users tend to follow a critical route described in the linguistic algorithm.

At this point, it is reasonable to present some search analyses that clearly illustrate some of the negative effects that Spanish Speakers commonly experience when they are consulting DB of the ISs dealt in here. They are divided in two groups: the first one displays the bridge between natural language and controlled vocabulary and, the second, shows the attempts users try to satisfy their information needs. In all cases, the searches were done on March, 2009.

Original Quest: *Bolsa Mexicana de valores*		**PROQUEST** ABI/Inform records	**EBSCO** Academic Search Premier records	**ELSEVIER** Science Direct records
Natural language	Mexican Bag of Values	4	1	1,475
Controlled Vocabulary	Mexican Stock Exchange	610	164	2,727

Table 1. Group 1

Original Quest: *Odontología*		**PROQUEST** Cambridge Medline records	**EBSCO** Academic Search Premier records	**Elsevier** Science Direct records
Nat. lang.	Odontology	1,893	1,394	13,563
Controlled Vocabulary	Dentistry	134,846	29,404	108,143

Table 2. Group 1

Original Quest: *Rana verde*		**PROQUEST** Biological Sciences records	**EBSCO** Academic Search Premier records	**Elsevier** Science Direct records
Nat. Lang.	Frog Green	11	1	19,574
Nat. Lang.	Green Frog	309	64	19,574
Controlled Vocabulary	Lithobates clamitans	1	52	13
Controlled Vocabulary	Rana Clamitans	262	40	550

Table 3. Group 1

Original Quest: *Protección Radiológica*		**Syntax and parts of the speech** Noun+Adj.	**PROQUEST Cambridge** Medline records	**EBSCO** Health Source: Nursing /Academic Edition records	**Elsevier** Science Direct records
First try	Radiology Protection	Noun+Noun	1,396	1	14,645
Second try	Radiologic Protection	Adj.+Noun	5	15,519	4,027
Third try	Radiological Protection	Adj.+Noun	1,336	67	16,152

Table 4. Group 2.

4. Infopragmatics

Infopragmatics is theoretically based on the pragmatics' aspects that Yule (1996) asserts: "Pragmatics is concerned with the study of meaning as communicated by a speaker (or writer) and interpreted by a listener (or reader)... … This type of study necessarily involves the interpretation of what people mean in a particular context and how the context influences what is said. It requires a consideration of how speakers organise what they want in accordance with who they're talking to, where, when and under what circumstances".

Infopragmatics is an efficient method to get pertinent information form ISs. It is based on a linguistic perspective aimed to those users, whose language is not English, paying particular focus, but not limited, to Spanish speakers. Infopragmatics may be of highly considerable help for those academicians who can not get pertinent information from commercial ISs such as the ones provided by Elsevier, Ebsco, ProQuest and the like and it consists of an uncertainty theorem, a linguistic algorithm and a linguistic storm (lingstorm).

4.1 Linguistic Algorithm

The linguistic algorithm emerges as a pragmatic solution (Ibarra, 2009) for IR. It is a formula that contains a series of symbols that prescribe the way in which a series of operations and reasoning should be done revealing the relationships among certain elements, which would be used to resolve a given problem (Beristáin, 1985).   This algorithm allows the users to consider appropriate steps to satisfy their information needs. It can be used as a guiding map to make users aware of the pair of routes involved within: the ideal and the critical.

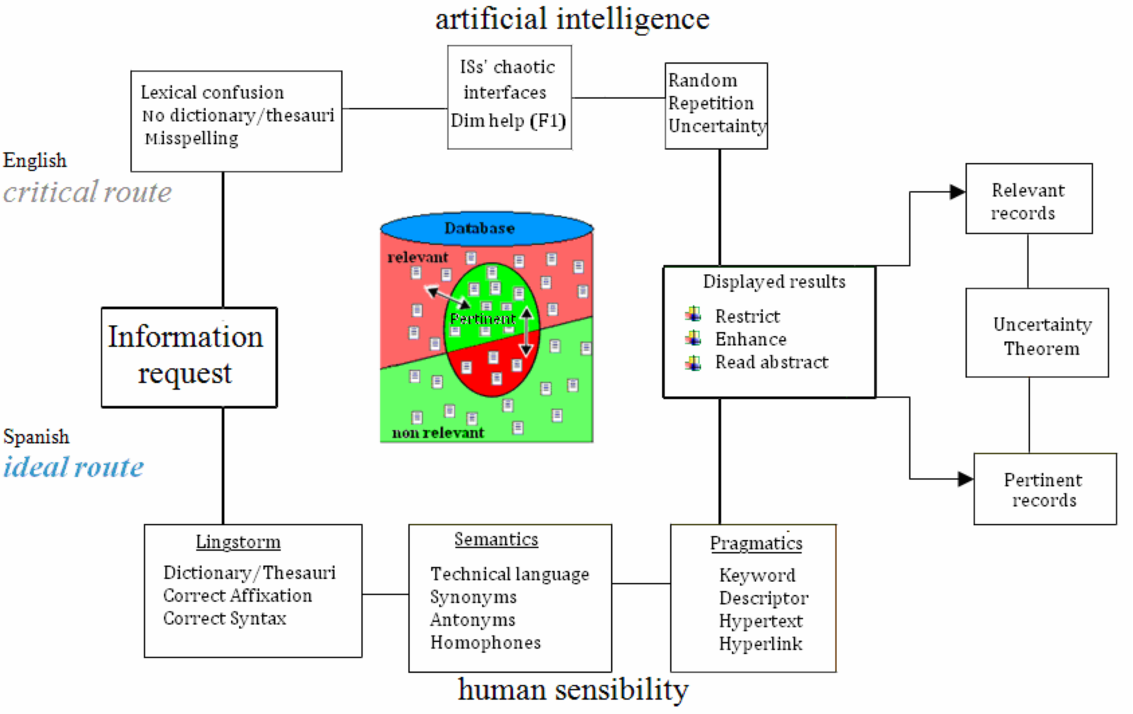


Fig. 2. Linguistic algorithm.

Furthermore, at the end of both routes, there is an extra element that substantially reduces the uncertainty state that may result in any of either direction.

4.2 The Uncertainty Theorem

When users can not get the information they need from a DB, both guided by librarians or by using the quick guides given by ISs providers, they wonder why, and there is not a chance to give them a clue to reveal what the problems were, since they had the chance to read *relevant* displayed records from a scientific ISs. For example, Ibarra (1999) presented an approach considering diverse linguistic theories from an ethnolinguistic perspective and several analyses based on 120 real information searches with similar frustrating results.

So far, ISs do not offer any aid to keep the users away from the resulted perplexity; a gap which is expected to be covered by a "series of subactivities and subgoals of factual information seeking", Allwood, J. and Haglund, B. (1991). These subactivities, subgoals and extra elements are resumed, grouped and designed in an uncertainty theorem as an efficient contrivance to steer users to step more confidently in their information goals:

I. If there is no information is due to...

- Users lack lexical abilities
- Syntax is wrong
- Used terms are not equivalent in meaning
- Context is no the appropriate
- There is nothing in the used data base

II. If the information does not fit is due to. . .

- Ineffective negotiation of meaning on technical terms
- Incorrect spelling
- Key words or descriptors are not such

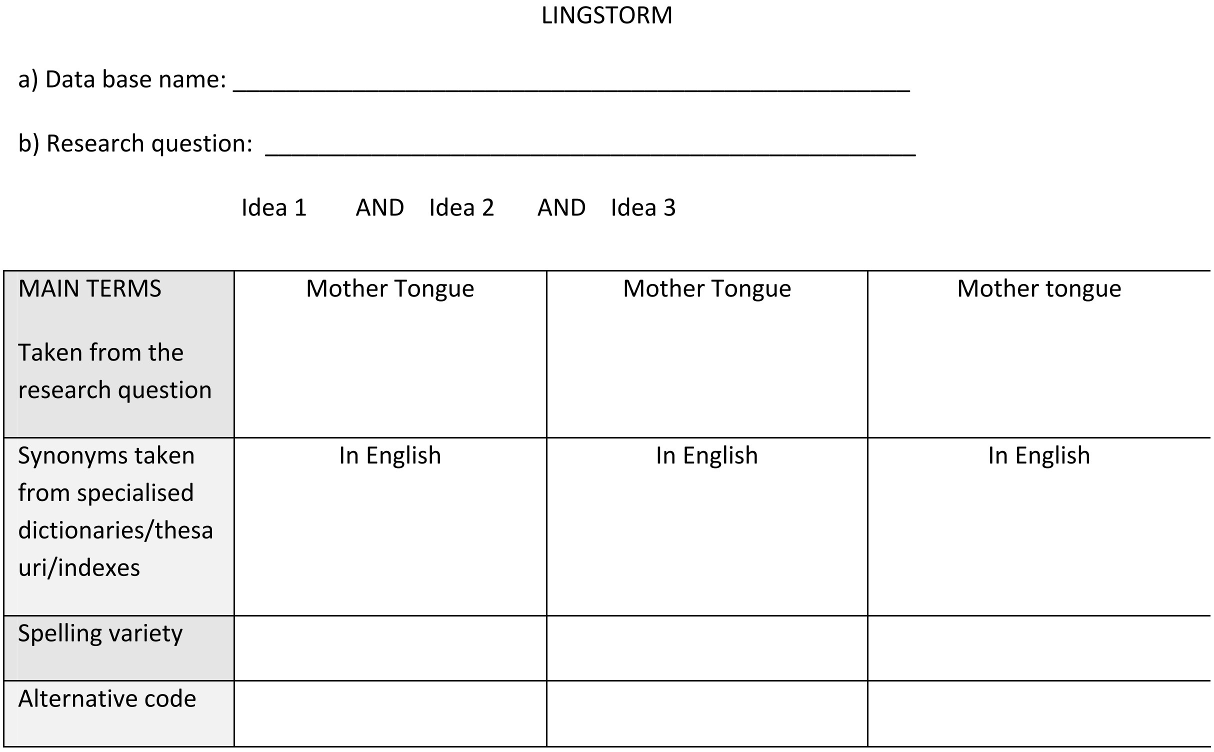
Liddy (2007) offers a useful table on the levels of Language Understanding, which Ibarra (2008) applies to a search of a Spanish native speaker search that results in a clear explanation of common failures: *Morphological*, bad spelling; *Lexical*, stating their needs not clearly; *Syntactic*, no grammar rules applied; *Semantic*, opposition between the natural language and controlled vocabulary;

Discourse, the way in which librarian and user interact to offer and receive ISs information respectively and, *Pragmatic*, the set of strategies that allow the users/librarians contrive their information needs.

The arguments from the theorem are supported by what Ellis (1990) states: “it is by no means easy to distinguish even questions, commands, and so on from statements by means of the few and jejune grammatical marks available such as word order, mood, and the like: though perhaps it has not been usual to dwell on the difficulties which this fact obviously rises. For how do we decide which is which? What are the limits and the definitions of each?"

4.3 Linguistic storm (*lingstorm*)

Specialised dictionary and thesaurus will allow the user to approach the *lingstorm*, “linguistic storm”, based on the suggestions of Hunter and Lodish (1989) and (Fig. 3), to be used in a similar way as a brainstorm works in a group or individual creativity exercise, a constant along the IR, needed to leap from natural language to controlled vocabulary. The correct affixation deciphers the differences among verbs, nouns and adverbs that commonly cause semantic breaks when users shift from Spanish to English; Syntax determines the results in noun phrases; technical language clinches users´ familiarity with controlled vocabulary; Synonyms enhance semantic horizons and include spelling varieties when dealing with non Latin characters; though antonyms and homophones take place during verbal interaction, they fathom a temporary aphasia that may be solved through an essential negotiation of meaning between the user and the librarian; key words and descriptors are easily found along the records´ reading and can be retrieved to set alternative search strategy; hypertext and hyperlinks are options that are attractive to follow as soon as they appear on the information records in order to avoid semantic unsettlement, though users tend to avoid them for preventing feeling lost if they lose the track of what they were looking for.



LINGSTORM

a) Data base name: ____

b) Research question: ____

Idea 1 AND Idea 2 AND Idea 3

MAIN TERMS Taken from the research question	Mother Tongue	Mother Tongue	Mother tongue
Synonyms taken from specialised dictionaries/thesauri/indexes	In English	In English	In English
Spelling variety			
Alternative code			

Fig. 3 Linguistic storm.

3.4 Statistical reasons and technical implementation of the *infopragmatics* in the National University of Mexico Library System

The technical implementation will demand observing the number of users, the collections and two users' scenarios: distance and On-site. As it was mentioned before, the UNAM received 46,000 - undergraduate and graduate – potential ISs users; according to the last statistical data, the UNAM's Library System, up to April, 2009, there is more than 2 million of documents, comprising books, thesis, ISs, serials, printed and electronic.

More specifically, the available 150 ISs – in English - comprise access to several millions of records, besides access to more than 1,000,000 of journals' issues. On the other hand, the information requests in ISs daily average is 20,000.

Considering the previous points, besides the level and kind of users, it is being designed a pair of brief workshops: On-site and distance. The first one consists of three parts: 15 minutes to explain the infopragmatics; 15 minutes of practice and; 15 for evaluation and feedback. It is necessary to mention that those taking this workshop and the distance one must know the academic databases dealt in this paper. The second workshop – distance- would be available on line and it is the users themselves who must set their times to profit from it, though there will be suggested steps and a feedback section via e-mail. Currently, there is a quick guide to "Retrieve pertinent information in English" in the Digital Library main web page <bidi.unam.mx>. The authors of this paper offer, to all those interested, to give a booklet containing the procedure of infopragmatics with the sole condition to provide them with feedback. Contact them.

From a qualitative point of view, infopragmatics workshops would be more effective in the On-site mode, nevertheless, due to the large quantity of users, and considering the three single steps of the algorithm, a quick guide was designed and entitled “Recupera información pertinente en inglés” (retrieve pertinent information in English) which is available at http://132.248.9.9:8084/infopragmatica/.

It is necessary to mention that this quick guide includes hyperlinks to electronic translators, specialised dictionaries and thesauri, which quality was evaluated and later, catalogued in the UNAM´s Digital Library.

There is a hyperlink that allows users to express their opinions on the utilisation of the quick guide at the bottom of the page. With the obtained results, that are stored in a database, the design would be improved.

……………………………………………………………………………

4. Conclusions

Information retrieval has a wide variety of problems that can only be solved by appropriate solutions, that is, if there are language disagreements, the clarifications must be given in language agreements. In this paper the common linguistic difficulties that affect Spanish speakers were identified as well as their corresponding explanations based on the presented evidence. Infopragmatics is an agile method that gives acute dynamism that allows ISs users to get pertinent information in opposition to the *relevant* ones that result from a critical route. The linguistic algorithm presented in here represents an accomplished tool that can be easily adapted to any language, included English, to serve as an information productive device.

……………………………………………………………………………………

References

ALLWOOD, J. and HAGLUND, B. Communicative Activity Analysis of a Wizard of Oz Experiment. Intenal Report PLUS (A Pragmatic Based Language Understanding System), ESPRIT P5254. 1991.

BAKHTINE, M. (Volochinov, V. N.) *Le marxisme et la philosophie du langage. Essai d'application de la méthode sociologique en linguistique*. [Paris:] Minuit. [traducido del ruso 1929.] (1977).

BERISTÁIN, H. Diccionario de retórica y poética. México, Porrúa. 508 p. 1985.

Dictionary of the history of Ideas. “The Grammarians”, Vol.2, p.664, 1st ed. 2003. http://etext.virginia.edu/cgi-local/DHI/dhi.cgi?id=dv2-74.

ELLIS, R. Instructed second language acquisition: Learning in the classroom / Rod Ellis. Oxford: Basil Blackwell. 230 p. 1990.

HUNTER, B. and LODISH, E. Online Searching in the Curriculum: A Teaching Guide for Library/Media Specialists and Teachers . USA: ABC-Clio, Santa Barbara, CA. (1989).

IBARRA, R. (1999). Aprovechamiento y optimización de los recursos tecnológicos en la búsqueda y recuperación de información en CD-ROM basados en estrategias lingüísticas. Tesis de maestría UACPyP – CELE UNAM. <http://pbidi.unam.mx/cgi-bin/ezpmysql.cgi?url=http://132.248.9.9:8080/tesdig/Procesados_1999/271153/Index.html> Acceso restringido al documento electrónico fuera de REDUNAM.

IBARRA, R. "Algunas grietas semánticas en la recuperación de información: una perspectiva deconstructiva para una solución pragmática". Primer Simposio Internacional Sobre Organización del Conocimiento: Bibliotecología y Terminología (27 – 29 agosto, 2007), CUIB - UNAM.

IDE, N. and VÉRONIS, J., (1998). "Introduction to the special issue on word sense disambiguation: the state of the art". Computational Linguistics, Vol. 24, No. 1, pp. 1-40.

JAFFE, J. (1988). "For Undergraduates: InfoTrac Magazine Index Plus or Wilsondisc with Reader's Guide and Humanities Index?", American Libraries, Vol. 19, No. 9, pp. 759-61.

KATSIRIKOU, A. AND SKIADAS C. H. "Chaos in the Library Environment", LIBRARY MANAGEMENT, VOL.22, NO. 6/7, PP. 278 – 287. 2001.

LIDDY, E. (2007). Whither Come the Words? Paper presented at the CENDI Subject Analysis and Retrieval Working Group Conference "Controlled Vocabulary and the Internet," September 29, [Power point presentation]. [On-line] <http://cendi.dtic.mil/presentations/liddy.PPT> [Consulted: 5 July, 2007].

Oxford Advanced Learner's Dictionary with Genie CD ROM. OUP., 2006.

STUBINZ. J. and WHIGHLY, S., Information Retrieval System Design for Very High Effectiveness. [On line]. (c. 1994)

VARONIS, E., & GASS, S. (1985). Miscommunication in native/nonnative conversation. Language in Society, 14, 327-343.

YULE, G. (1996) Pragmatics. Oxford University Press. P.138.