A step forward to hypertext

**Adán Cassán, Sergi Cervell, Mireia Colom, Rafael Marín,
Josep M. Merenciano, Gema Pérez, Lluís Valentín**

Department of Computational Linguistics, Planeta Actimedia
C/ Aribau, 198, 5ª planta, 08036, Barcelona, Spain
{rmarin, jmmerenciano, gperez, lvalentin}@planeta-actimedia.es

Abstract

In this paper, after a critical review of how hypertext has been understood over the past few years, we claim against the distinction between total and partial hypertext, and we provide a brief description of a dynamic system that allows the automatic highlighting of those textual elements related to a certain topic. The outcome of our approach is ESQUITX, an automatic highlighter based on different filters, particularly those referring to topic information. The general process can be summarized as follows: once the text is lemmatized, by means of our Spanish tagger, a collection of filters is applied, and only the resulting lemma forms are highlighted.

1. Introduction

In 1945, Vannevar Bush presented the Memex project, a system that, although it was never implemented, it was the starting point of which we know as hypertext: "a provision whereby any item may be caused at will to select immediately and automatically to another" (Bush 1945).

In 1963, Douglas Engelbart developed the first hypertextual system: NLS/Augment, designed as an experimental device for the scientific community to file and share their articles (Engelbart 1963).

In 1965, Ted Nelson coined the term "hypertext" to describe a "body of written or pictorial material interconnected in a complex way that it could not be conveniently represented on paper" (Nelson 1965). His project Xanadu is, in a way, the precedent of which we now know as hypertext, that is, an information system where data objects connect to each other through links, on a creative and disorganised basis, without following any pre-established structures.

In 1987, Apple Computer used the HyperCard tool as the main sale argument for the Macintosh. Created by Bill Atkinson, it was a system of personal, modifiable and programmable organization, essentially based on hypertext.

In 1989, Tim Berners-Lee and Robert Caillau, while at CERN, developed HTML (HyperText Markup Language), a standard language that specifies the format or structure of documents (indicated by markup tags) for retrieval across the Internet, together with the help of HTTP (HyperText Transfer Protocol). HTML/HTTP document servers were called WebServers, and it raised the name of WorldWideWeb to the assembly of all.

In 1991, Marc Andreessen and Eric Bina, also members of CERN, developed Mosaic, the first popular web browser before Netscape, widely used across the world. That meant only a first step on the way: the web began to grow exponentially in the twinkling of an eye, as well as the number of users. The "global village" started to become real (Horney 1991; Sutherland 1997; Campás 2001).

Nowadays, hyperlinks turned into a necessary feature rather than an added value. Nowadays, it is just unacceptable a static multimedia product where hyperlinks simply do not exist. This change, though, must be analysed in terms of quantitativeness: the amount of accessible information as well as the number of links contained in a document have increased, but neither the type of information linked (text, pictures, sound, video, etc.) nor the management have changed that much. (Streitz et al. 1990; Landow 1997).

Perhaps we can find the most explicit qualitative advance in the creation of hyperlinks: procedures such as lemmatization allow to automatically assign any word of a text (appearing in singular, plural or past perfect tense) to a lemma; and this one to a document (generally an encyclopaedic article or a dictionary entry, in the case of multimedia encyclopaedias). The result is the so-called **total hypertext**.

The project we present leaves aside, in a way, the traditional hypertext, to go a little bit further. Our proposal is a dynamic vision of documents, a system that can be customized by the users according to their needs. To be exact, the outcome of our approach is ESQUITX: an automatic highlighting system based on topic information.

Although the project was designed for multimedia encyclopaedias, it is worth to note that it can be implemented on any hypertextual-based system, such as the web.

2. A step forward to hypertext

Links are traditionally represented by means of highlighting those elements in a text having information associated to them (that is, the link itself). At the present time, hypertextual systems are differently conceived, thanks to the implementation of total hypertext: if we have a system where each word in a text has a link, the highlighting looses its original purpose (that is, to guide through words in a text having a link).

As a result, we may reconsider the logic and utility of highlighting: if highlighting does not work as a *navigation wizard* (indicating the presence of links in a text) then it must have new reasons to be there, must follow other purposes, such as, for instance, to help to *conceptual navigation* or *text comprehension*.

The final goal of highlighting, whatever it may be, needs to be carefully considered. On the one hand, we can include in a text (using SGML markup tags, for instance) all the information required for the browser to manage it as hyperlinks. On the other hand, we can design tools and mechanisms that are able to extract only the necessary information for the same purpose directly from plain text. In the latter case, the manually markup task (with static results) would disappear, making way for the dynamism of texts, as we will see.

The manual (or static) solution is that offered by conventional hypertext. Total hypertext automatically generated (by means of lematization, for instance), though involved in automatic processes, also gives static results. The static solution is also used for highlighting at the so-called selective hipertext (words highlighted because they contain a manually reviewed link); or conceptual hyperlinks (generally covering more than a word, and sending to documents that are impossible to obtain by means of lemmatization or similar mechanisms).

Our concerning is to find a global, automatic-based solution, for both the hyperlinks and the highlighting (for which we also think of a dynamic solution). To be exact, our goal is that the user of a document may be able to decide which words should be highlighted. It contributes to many benefits: 1) production costs would be substantially lower thanks to the automatization of highlighting tasks and the reuse of texts; 2) the decision-taking processes with regard to what is going to be highlighted would disappear; 3) last (but not least) the user of the text would be the one who decides and models the highlighting according to his or her needs, and not to the needs of the publisher.

To sum up, we find a global, automatic and dynamic solution: a hypertextual system where all the processes involved are automatic and where total and customized hypertext do coexist, with user-defined support to text comprehension.

With this perspective in mind, the highlighting as a help to text comprehension could be conceived as a system combining automatic summarisation and keyword identification. That is, the system automatically summarises the text and tries to reproduce its semantics highlighting certain words, so the user can take, reading through it, a general idea of the meaning or main topic of the text.

The outcome of our approach is not so ambitious. It is, essentially, a system standing between text comprehension and "conceptual" navigation support.

3. Our proposal

As a result of our approach we have designed and implemented ESQUITX, a system that automatically highlights words in a text based on topic information.

Roughly speaking, ESQUITX proceeds as follows: given a text and a topic, a morphological analyzer produces a set of valid candidate lemmas which are linked to nodes in a tree-like topic structure called ATEM. The set of candidates then undergo a filtering process. The resulting lemmas will be highlighted.

As we shall see, the overall computational complexity is low, due to the coupling of ATEM and our POS tagger (henceforth DICO).

3.1. The system

A suitable operation of ESQUITX is based on: 1) a Spanish tagger; 2) a structured topic tree; 3) a topic assignment of each lemma; and 4) a reference topic. Let us see in detail these four requirements.

3.1.1. The POS tagger

From a given input text, we must be able to recognize most words as we can and associate them to encyclopaedic entries. For this purpose, we have a system that allows connecting every word in a text to its related lemma. For instance, our system recognizes *perros* ('dogs') as M(asculin) P(lural) C(ommon) N(oun) and, therefore, associates the word to the lemma *perro* ('dog'). In a similar way, the forms *F. García Lorca*, *García Lorca* and even *Lorca* are recognized and linked to its reference form *Federico García Lorca* (cf. Casanova *et al.* 2001).

Our morphological analysis system includes, as well as the usual lemmatization and categorization modules, some especial features, like a module for derivative morphology (that helps us to catch, for example, augmentatives, diminutives, *-mente* ended adverbs and unknown words) and another module for morphological disambiguation, which accuracy is now over 95%.

In addition to this, it is worth to point that the base of the system, DICO, is connected to an encyclopaedic lexicon, so, besides lemmatising, we can suggest hyperlinks to entries in the encyclopaedia.

3.1.2. The topic structure

The ATEM topic structure contains about 800 nodes, organized in a hierarchy of seven levels. The nodes have a transitive relation, without denoting classes, but a topic area, mainly defined according to pragmatic criteria.

In Figure 1 there is a demonstrative example: the node *Epistemología* ('Epistemology') has, as sons, *Filosofía de la ciencia* ('Philosophy of Science') and *Filosofía de la mente* ('Philosophy of Mind') and, as father, *Filosofía* ('Philosophy'), which in turn, is a son of *Humanidades y Arte* ('Humanities and Arts'), one of the four first level nodes of ATEM.

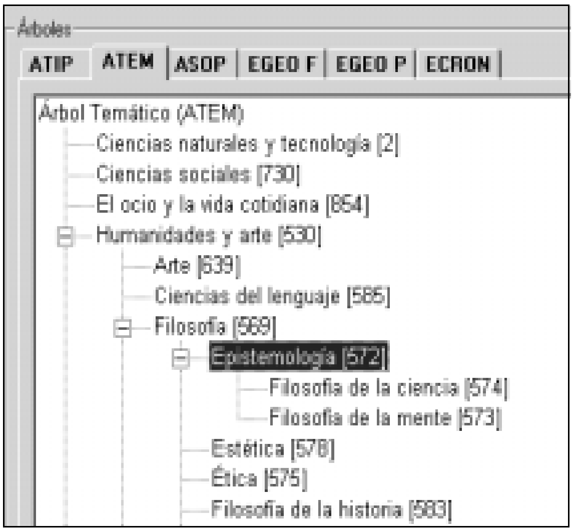

Árboles
ATIP
ATEM
ASOP
EGEO F
EGEO P
ECRON
Árbol Temático (ATEM)
Ciencias naturales y tecnología [2]
Ciencias sociales [730]
El ocio y la vida cotidiana [854]
Humanidades y arte [530]
Arte [639]
Ciencias del lenguaje [585]
Filosofía [569]
Epistemología [572]
Filosofía de la ciencia [574]
Filosofía de la mente [573]
Estética [578]
Ética [575]
Filosofía de la historia [583]

Figure 1. An example of the ATEM topic tree.

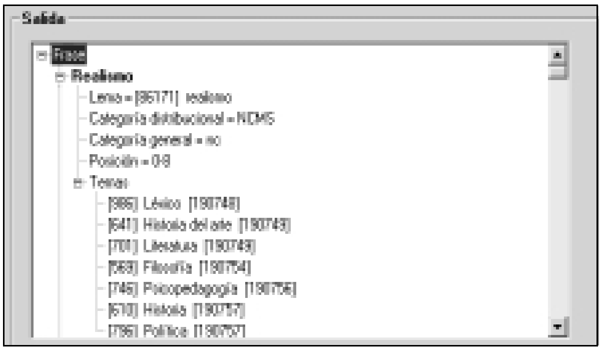
The ATEM's structure causes one of the main features of our system; thus, if we had just a plain topic list (as is usually done with subject fields in publishing works) we could not carry on most of the processes we can afford now.

3.1.3. The topic assignment

Each lemma is linked to all its related word senses existing in a Knowledge Base (from here on, BDCon), and each word sense is associated, at least, to one topic in the ATEM tree.

Hence, the topic assignment does not really take place between lemmas and topics, but between word senses and topics. However, as the tagger cannot work directly with word senses, ESQUITX joins the union of topics assigned to all the word senses related to a lemma, so it is, by this way, able to assign right topics to almost every lemma. An interesting option is asking ESQUITX to choose some groups of word senses before rating the union of topics. Thus, we could just be interested in main entries or very frequent entries of a given word.

Let us see a particular example. As can be observed in Figure 2, the noun *realismo* ('realism') has five word senses in BDCon. Three of this word senses (IDs 190748, 190754 and 190756) are linked to a single topic, respectively, to *Léxico* ('Lexical', i.e. without an specific topic), *Filosofía* ('Philosophy') and *Psicopedagogía* ('Psicopedagogy'); and two of them are linked to a couple of topics: ID 190749 is related to *Historia del Arte* ('Art History') and *Literatura* ('Literature'), whereas ID 190757 is related to *Historia* ('History') and *Política* ('Politics').



Salida

Frase
Realismo
Lema = [86171] realismo
Categoría distribucional = NCMS
Categoría general = nc
Posición = 0-8
Temas
[986] Léxico [190748]
[641] Historia del arte [190749]
[701] Literatura [190749]
[569] Filosofía [190754]
[746] Psicopedagogía [190756]
[610] Historia [190757]
[786] Política [190757]

Figure 2. An example of tagging and topic assignment.

There is also included some data about morphological analysis: the word *Realismo* is a M(asculin) S(ingular) C(ommon) N(oun), associated to the lemma *realismo*, ID 86171.

3.1.4. The reference topic

To make ESQUITX operative is strictly necessary to have a reference topic or subject field. So, for instance, if we want to highlight the text of the entry *célula* ('cell'), this entry requires a topic. If it is *Biología* ('Biology') all the words in the text which lemmas are related to Biology will be highlighted.

At this point, our proposal clearly shows its own dynamism: who brings ESQUITX a reference topic is the system user, that is, the reader of the file. It does not mean that the publisher or the administrator of the file could not suggest a default topic. Joining both possibilities allows the system to be flexible and easy to change.

Thus, for instance, the entry *célula* could have, by default reference topic *Biología*, and the user could change it for a more specific one, like *Biología Molecular* ('Molecular Biology'), or maybe for a more generic one, like *Ciencias de la Vida* ('Life Sciences'), or even a completely different one, like *Derecho* ('Law') or *Deportes* ('Sports').

This dynamism brings an additional feature in front of the inherent stillness of the usual hypertext systems, and especially regarding to those manually highlighted.

3.2. ESQUITX

ESQUITX can be seen as a filtering system that acts upon a sequence of lemmas enriched with topic information. Lemmas are obtained from the morphological information of the input text, while the corresponding topics come from the relation of each word sense to a node of the topic tree ATEM, as mentioned above.

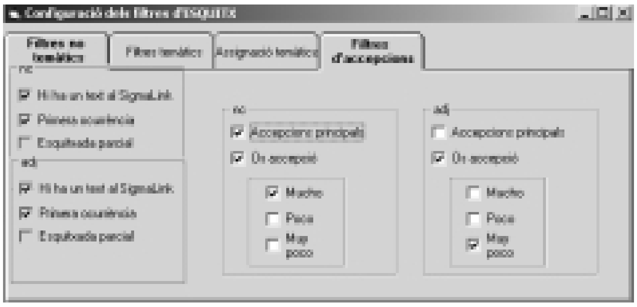
Filters are independent and cummulative. They are cummulative as we can consider every filter as a lemma selector, and only those lemmas which succeed in a filtering come into consideration for the following one. The independence of the filters implies that the ordering of their execution does not affect the final result. This is what allows ESQUITX to decide the runnig of the process in computing terms exclusively.

Once all the filters have been applied the final output comes to a last selection, which is not exactly a filter as it is not independent (it must come in the last place). It sets a number (absolute, percentage, fixed or random) of successful highlighted words in the processed text.

Two main kind of filters have been used: topic and non-topic filters. Among non-topic filters, the following ones should be pointed out:

- **Filter by category.** ESQUITX has been designed to only consider common nouns, adjectives and entity names. It is the user who decides which categories are to be implied.
- **Filter by link.** Only those words with an active link to an encyclopaedic entry succeed.
- **Filter by order of occurrence.** Only the first occurrence of each lemma succeed.
- **Filter by reference form.** It is applied only to entity names. The ones which are recognised as reference forms of the name succeed (Casanova *et al*. 2001).

Figure 3 shows some of the options related to non-topic filters:



Configuració dels filtres d'ESQUITX

Filtres no temàtics | Filtres temàtics | Assignació temàtica | Filtres d'accepcions

nc
☑ Hi ha un text al SignaLink
☑ Primera ocurrència
☐ Esquitxada parcial

adj
☑ Hi ha un text al SignaLink
☑ Primera ocurrència
☐ Esquitxada parcial

nc
☑ Accepcions principals
☑ Un accepció
☑ Mucho
☐ Poco
☐ Muy poco

adj
☐ Accepcions principals
☑ Un accepció
☐ Mucho
☐ Poco
☑ Muy poco

Figure 3. An example of non-topic filters.

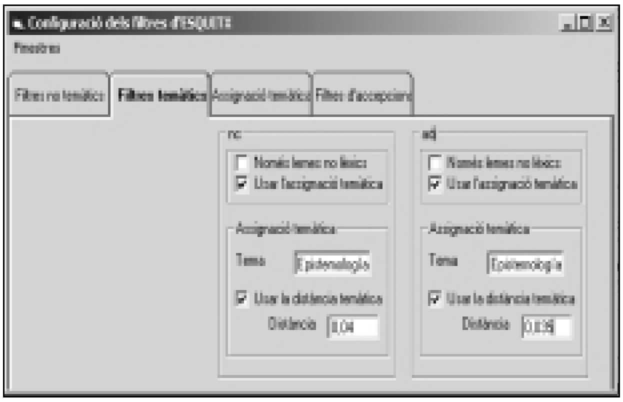
As for those filters based on topic information the most representative are:

- **Reference topic filter.** Only those words in the text directly associated to the topic of the document succeed.
- **Narrowing filter.** It is only succeeded by those words directly associated to the topic of the document or to those related topics in lower levels of the topic tree structure.
- **Broadering filter.** It is only succeeded by those words directly associated to the topic of the document or to those related topics in higher levels of the topic tree structure.
- **Topic-related filter**. The success criteria is set by those ATEM nodes having a specific distance (or lower) to the reference topic.

Bearing in mind that the topic information is contained in a topic tree structure, it is important to stress the concept of topic distance between nodes.

The concept of distance used by ESQUITX allow the automatic processing to capture the common sense relation among topics: difference, coincidence and proximity. A proper use of this distance must subsume the previously commented topic filtering.

The following Figure presents some of the possible topic filtering configurations:



nc

☐ Només lemes no lèxics
☑ Usar l'assignació temàtica

Assignació temàtica

Tema: Epistemologia
☑ Usar la distància temàtica
Distància: 0,04

adj

☐ Només lemes no lèxics
☑ Usar l'assignació temàtica

Assignació temàtica

Tema: Epistemologia
☑ Usar la distància temàtica
Distància: 0,035

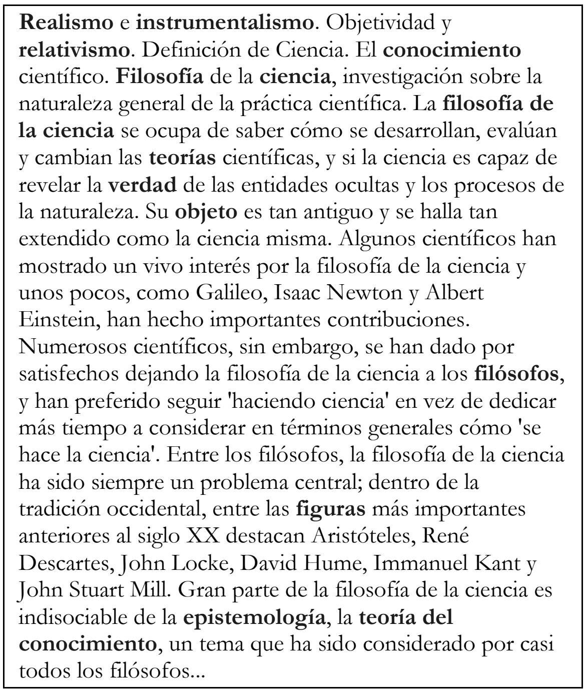
Figure 4. An example of topic filters.

It is important to remember that the interface allows to use any combination of all the filtering options previously mentioned. Besides, ESQUITX can also be configurated with some parameters.

It is worthwhile to note that such a simple mechanism, if based on a powerful knowledge base, provides really interesting results, as it can be seen in the following example:

3.3. Two examples

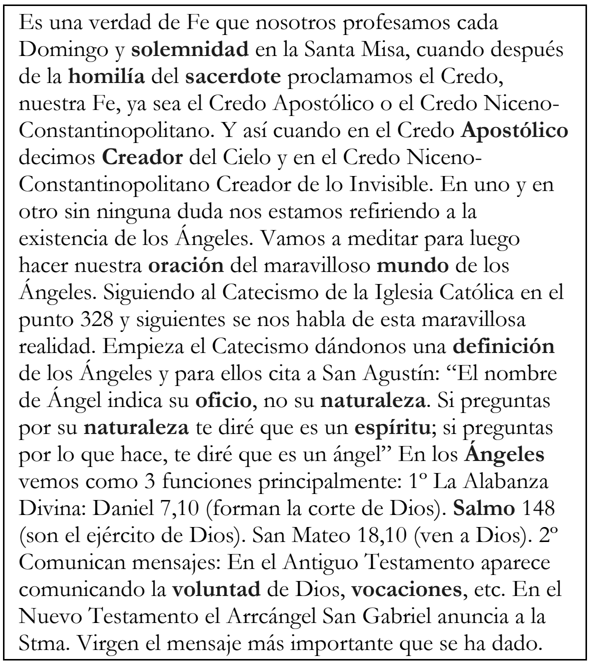
In order to highlight a text, as the one included in Figure 5, the following configuration is defined: taken *Epistemología* as a reference topic, highlight those common nouns assigned to *Epistemología* and his sons. Here is the result:



Realismo e **instrumentalismo**. Objetividad y **relativismo**. Definición de Ciencia. El **conocimiento** científico. **Filosofía** de la **ciencia**, investigación sobre la naturaleza general de la práctica científica. La **filosofía de la ciencia** se ocupa de saber cómo se desarrollan, evalúan y cambian las **teorías** científicas, y si la ciencia es capaz de revelar la **verdad** de las entidades ocultas y los procesos de la naturaleza. Su **objeto** es tan antiguo y se halla tan extendido como la ciencia misma. Algunos científicos han mostrado un vivo interés por la filosofía de la ciencia y unos pocos, como Galileo, Isaac Newton y Albert Einstein, han hecho importantes contribuciones. Numerosos científicos, sin embargo, se han dado por satisfechos dejando la filosofía de la ciencia a los **filósofos**, y han preferido seguir 'haciendo ciencia' en vez de dedicar más tiempo a considerar en términos generales cómo 'se hace la ciencia'. Entre los filósofos, la filosofía de la ciencia ha sido siempre un problema central; dentro de la tradición occidental, entre las **figuras** más importantes anteriores al siglo XX destacan Aristóteles, René Descartes, John Locke, David Hume, Immanuel Kant y John Stuart Mill. Gran parte de la filosofía de la ciencia es indisociable de la **epistemología**, la **teoría del conocimiento**, un tema que ha sido considerado por casi todos los filósofos...

Figure 5. An example of a text automatically highlighted

The reference topic of the following text is *Religión* ('Religion'), the collection of filters applied remaining invariable:



Es una verdad de Fe que nosotros profesamos cada Domingo y **solemnidad** en la Santa Misa, cuando después de la **homilía** del **sacerdote** proclamamos el Credo, nuestra Fe, ya sea el Credo Apostólico o el Credo Niceno-Constantinopolitano. Y así cuando en el Credo **Apostólico** decimos **Creador** del Cielo y en el Credo Niceno-Constantinopolitano Creador de lo Invisible. En uno y en otro sin ninguna duda nos estamos refiriendo a la existencia de los Ángeles. Vamos a meditar para luego hacer nuestra **oración** del maravilloso **mundo** de los Ángeles. Siguiendo al Catecismo de la Iglesia Católica en el punto 328 y siguientes se nos habla de esta maravillosa realidad. Empieza el Catecismo dándonos una **definición** de los Ángeles y para ellos cita a San Agustín: "El nombre de Ángel indica su **oficio**, no su **naturaleza**. Si preguntas por su **naturaleza** te diré que es un **espíritu**; si preguntas por lo que hace, te diré que es un ángel" En los **Ángeles** vemos como 3 funciones principalmente: 1º La Alabanza Divina: Daniel 7,10 (forman la corte de Dios). **Salmo** 148 (son el ejército de Dios). San Mateo 18,10 (ven a Dios). 2º Comunican mensajes: En el Antiguo Testamento aparece comunicando la **voluntad** de Dios, **vocaciones**, etc. En el Nuevo Testamento el Arrcángel San Gabriel anuncia a la Stma. Virgen el mensaje más importante que se ha dado.

Figure 6. Another example of a text automatically highlighted

As can be observed, there are some errors. Thus, for example, in Figure 5, the first appearance of *Ciencia* has not been highlighted, because the tagging has categorized it as a proper noun.

4. Conclusions and future work

As we have seen, *total hypertext* seems to challenge the need for partial word highlighting. In order to preserve it, we have to extend the meaning of highlighted words in a text, from merely indicating a link to a semantically richer role.

We have addressed this problem computationally by means of ESQUITX, a low computational cost system capable of identifying all the words in a text related to a given topic. We have proven that a reduced set of clear and systematic criteria are enough to obtain excellent results.

Even if ESQUITX has been designed with multimedia encyclopaedia in mind, it should be seen as a general highlighting tool for Spanish texts. Furthermore, due to its simplicity and efficiency, ESQUITX could be run online or from a CDROM.

Nonetheless, just as any piece of software, ESQUITX can be improved in many ways. Our research is currently focused on a weighted topic linkage process, which could also be used to perform other linguistic tasks such as word sense disambiguation, information retrieval or text classification.

In our opinion, ESQUITX sheds new light upon hyperlinks' meaning, and suggests that topic information could play a significant role in a great number of NLP applications.

5. References

Bush, V. (1945). As we may think. *Atlantic Monthly*, 176(1): 101-108.

Campàs, J. (2001). Hipertext. Tècnica d'escriptura i creació. In M. A. Martín (coord.). *Llengua catalana IV: les tecnologies del llenguage*, Barcelona: UOC.

Casanova, D., Lloré, X., Marín, R., Merenciano, J. M., Pérez, G. y Trotzig, D. (2001). ANTRO: Un sistema de reconocimiento y gestión de antropónimos. *Actas del XVII Congreso de la SEPLN*, 311-312. Also in http://www.planeta-actimedia.es/esp/banco/ling.htm.

Engelbart, D. C. (1963). A Conceptual Framework for the Augmentation of Man's Intellect. In P. Howerton (ed.), *Vistas in Information Handling*, 1, Washington DC: Spartan Books.

Horney, M. A. (1991). Uses of Hypertext. *Journal of Computing in Higuer Education*, 2: 44-65.

Landow, G. P. (comp.) (1997). *Teoría del hipertexto*. Barcelona: Paidós.

Nelson, T. (1965). A File Structure for the Complex, the Changing and the Indeterminate. *20th ACM National Conference*.

Streitz, N., Rizk, A. y André, J. (eds.) (1990*). Hypertext: concepts, systems and applications*. Cambridge: Cambridge University Press.

Sutherland, K. (ed.) (1997). *Electronic Text*. Oxford: Clarendon Press.