Procesamiento del Lenguaje Natural, nº39 (2007), pp. 139-146 recibido 18-05-2007; aceptado 22-06-2007

Towards Quantitative Concept Analysis

Rogelio Nazar
rogelio.nazar@upf.edu

Jorge Vivaldi
jorge.vivaldi@upf.edu

Leo Wanner
leo.wanner@upf.edu

Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Pl. de la Mercè 10-12
08002 Barcelona

ICREA and Dept. de Tecnologías de la Información y las Comunicaciones
Universitat Pompeu Fabra
Passeig de Circumval·lació 8
08003 Barcelona

Abstract: In this paper, we present an approach to the automatic extraction of conceptual structures from unorganized collections of documents using large scale lexical regularities in text. The technique maps a term to a constellation of other terms that captures the essential meaning of the term in question. The methodology is language independent, it involves an exploration of a document collection in which the initial term occurs (e.g., the collection returned by a search engine when being queried with this term) and the building of a network in which each node is assigned to a term. The weights of the connections between nodes are strengthened each time the terms that these nodes represent appear together in a context of a predefined length. Possible applications are automatic concept map generation, terminology extraction, term retrieval, term translation, term localization, etc. The system is currently under development although preliminary experiments show promising results.

Keywords: Corpus Linguistics; Concept Map Generation; Term Retrieval

Resumen: En este trabajo presentamos una aproximación a la extracción automática de estructuras conceptuales a partir de colecciones desordenadas de documentos, aprovechando regularidades léxicas a gran escala en los textos. Es una técnica para asociar un término con una constelación de otros términos que refleje lo esencial del significado. La metodología es independiente de la lengua. Se explora una colección de documentos donde el término inicial aparece (como la colección que devuelve un motor de búsqueda con esa interrogación) y se construye una red en la que cada nodo es asignado a un término. La ponderación de las conexiones entre nodos se incrementa cuando los términos que representan aparecen juntos en un contexto de extensión predefinida. Posibles aplicaciones son la generación automática de mapas conceptuales, la extracción de terminología, la recuperación de términos, su traducción, localización, etc. El sistema se encuentra actualmente en desarrollo, sin embargo experimentos preliminares muestran resultados prometedores.

Palabras clave: Lingüística de corpus; Generación de mapas conceptuales; Recuperación de términos

1 Introduction

In this paper, we describe an technique that, starting from a query term provided by the user and a document collection, generates a network of terms conceptually related to such query term. The resulting network is assumed to reflect the most pertinent information found in the collection in relation to the query term.

We call such networks *concept maps* since, in accordance with the relational paradigm of lexical memory (see, e.g., Miller, 1995), we presuppose that the meaning of a term (i.e., a *concept*) is given by all relevant relations that hold between this term and other terms – with the totality of these relations resulting in what is commonly known as a *map*.

The generation of the conceptual maps in our algorithm is guided by quantitative means.

ISSN: 1135-5948

More precisely, it is based on the most recurrent combination patterns among terms in a given document collection.[1]

The work presented here differs in both its theoretical assumption and its objective from the ontology generation field (cf. Buitelaar et al., 2005 for an overview). Our work is not ontological because we are not interested in what something IS. Rather, we are interested in what people usually SAY about something. We extract a synthesis of people's perception in reference to a topic from a whole set of documents rather than information from individual sources. Furthermore, we analyze how concepts evolve in real time as result of massive amounts of statements disseminated via the web. This is knowledge whose evolution is based on the same mechanism as self-organized complex systems.

Our intuition is that this is also how common knowledge is being developed. For instance, common knowledge tells us that alchemists wanted to transmute metals into gold. And it turns out that the word *alchemist* has a strong statistical association with words such as *transmute* and trigrams such as *metals into gold*. The present work is therefore less related to Artificial Intelligence (AI) than it is to linguistics. In fact, it is an example of "artificial"-AI, because it relies on social networks and the unconscious collaborative work of a collective of authors.

The remainder of the paper is structured as follows. In the next section, we present the hypothesis underlying our work. Section 3 outlines the methodology we adopt, and Section 4 illustrates our proposal by a couple of examples. In Section 5, a short overview of the related work is given, before in Section 6 some conclusions and directions for future work are drawn.

2 Hypothesis

The question underlying our work is: How is it possible to distinguish relevant information from irrelevant information with respect to a given specific term? In particular, how is it possible to make this distinction by means of a formal prediction instead of subjective or arbitrary judgment? From our point of view, this is possible through the study of large-scale regularities in the lexical organization of the discourse.

Adopting the relational paradigm of the structure of lexical memory (see above) and assuming that the recurrent context of a term reflects the comprehension of this term by the speakers, we draw upon frequency distribution as the decisive means for the construction of a conceptual map. Further theoretical evidence supports the idea of systematic redundancy in the surrounding context of a term. Following Eco (1981) we assume that textual devices such as appositions, paraphrases or coreferences let the writer mention attributes of a referent without compromising assumptions on the knowledge of the reader. The writer has a *model reader*, an idea about what the reader may or may not already know. Consider an example:

(1) *This is an image of Napoleon Bonaparte, Emperor of the French and King of Italy, looking unamused at...*

(1) shows the use of an apposition that is equivalent to the plain proposition:

(2) *Napoleon Bonaparte was the Emperor of the French and King of Italy.*

There are myriads of utterances about Napoleon, all different at the surface, but there is also a space of convergence, which we perceive as patterns of recurrent key terms – including those that appear in (2). Thus, in the list of most frequent terms that occur on May 3rd, 2007 in the web in the context of Napoleon we encounter, among others: *emperor, France*, *Bonaparte*, *invasion*, *Russia*, *king*, *Italy*, … *French*, …

These units roughly follow a Zipfean distribution: only a relative small number of them show a significant cooccurrence and this is why we can apply statistics to grasp them.

3 Algorithm

In this paper we propose an algorithm that accepts a term as input and uses it as query for an off-the-shelf search engine. From the document list retrieved by the engine, a parameterizable number of documents is downloaded. From these documents, the algorithm builds a conceptual map for that query. A vocabulary selection is performed and only the chosen units are considered during the

[1] Henceforth, we use the terms "term" and "lexical unit" as equivalents in this paper.

map construction. The overall process consists of five major steps:

A. Extraction of the contexts of the occurrence of the query term in the document collection. The contexts consist of a parameterizable number of words (15 by default) to the left and to the right of the term (we are not interested in sentence boundary detection since semantic association transcends it).

B. Compilation of an index from the extracted contexts. In addition to single tokens, the index includes a list of bigrams and trigrams, henceforth, *n*-grams (n = 2, 3). From this index, items that begin or end with a member of a stopword-list are excluded. This stoplist contains punctuation marks, hyphens, brackets, functional (i.e., closed class) words and optionally numbers. It was extracted from the first hundred positions in the list of word frequencies of nine languages obtained from Quasthoff (et al. 2005).

C. Merge of different word forms considered to be similar.[2] This procedure identifies inflectional variations (as, e.g., *animals* and *animal*) and reduces them to the same word (namely, the most frequent form among the variations) computing a Dice similarity coefficient with trigrams of characters as features, only if both variants have the first trigram in common.

D. Elimination of irrelevant terms from the index. Further reduction of the vocabulary is executed by removing terms and *n*-grams of a frequency below a predefined threshold (usually 4 or 5). Also, terms that appear in only one document are eliminated. The rest is filtered using statistical measures such as Mutual Information (MI), t-score, and chi-square. The threshold score for the association is another parameter, but by default it is automatically adjusted to meet the best conditions. The expected probability of the occurrence of words has been extracted from Quasthoff et al. (2005)'s model, but not with data for low frequency words (f<6). As a result, if a term is not listed there, it is treated as if it was, but with the minimum frequency.

E. Construction of the conceptual map. The algorithm reads all contexts of the query term and if the terms encountered in these contexts are in the selected vocabulary, each of them is assigned to a unique node in the network. The connections between these nodes are strengthen each time the terms associated with the nodes appear in a context. Every time an edge is stimulated, the rest is weakened. As the learning progresses, the weight of the nodes is weakened if they were assigned a particular term at the beginning but found no significant connections with neighbors afterwards. At the end of the learning process, the most interconnected nodes are key terms related to the meaning(s) of the query term. The nodes also have references to the original documents and contexts where their terms occur. The final number of nodes is determined by an initial parameter, and several prunes may be conducted to reduce nodes until this number is reached.

[2] Note that we do not use lemmatization and POS-tagging. We were interested in measuring accuracy without this type of resources.

4 Preliminary Results

A few experiments with this algorithm showed that it performs as expected. Currently, we are about to carry out a more extensive and formal evaluation that will allow us to provide exact figures of accuracy.

To give the reader an overview of the algorithm's potential and applicability, we briefly illustrate its performance in a few applications.

4.1 Concept Mapping

The most basic application is to obtain a map of terms conceptually related to the given query term. The terms captured in the network of the query term DUCKBILL PLATYPUS (Figure 1) are precisely its salient attributes: *ornithorhyncgus anatinus; fur; swimming animal*; *unique species; mammal; lay eggs; spiny anteaters;* etc.

Figure 1: Network for DUCKBILL PLATYPUS

Note that the network contains most of the terms needed for the generation of the lexicographic definition for DUCKBILL PLATYPUS:

(3) **Duckbill platypus**: *ornithorhyncgus anatinus, furred swimming animal, unique species of mammals that lay eggs, along with the spiny anteaters.*

4.2 Term Disambiguation

Given a polysemous term as a query, the network shows clustering effects for each sense. For instance, with the Spanish word HENO (*hay*), different clusters are visible. Figure 2 shows a fragment of this network.

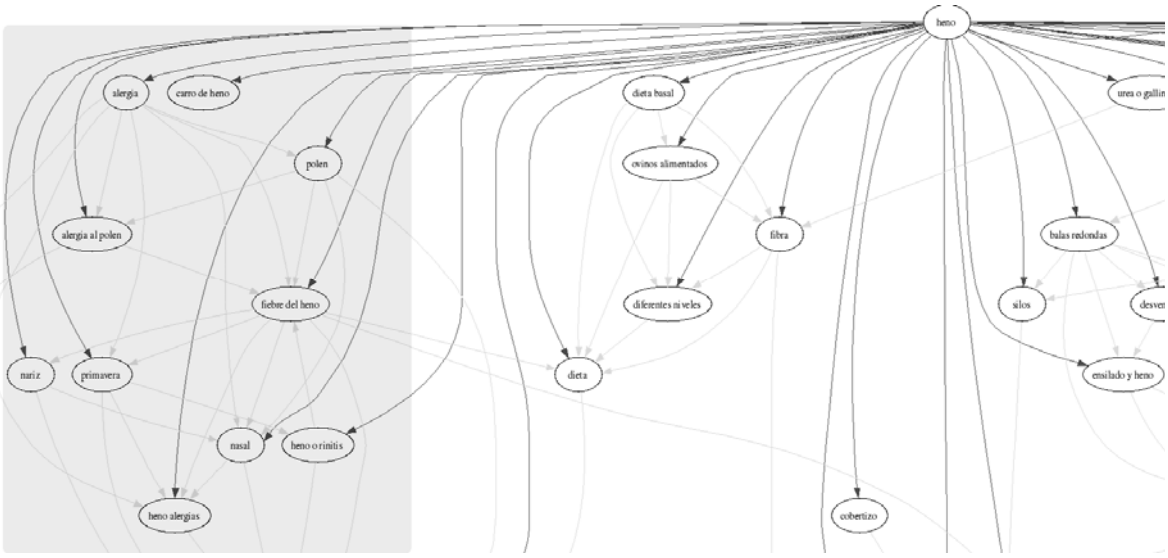

Figure 2: Network for HENO

At the left hand side there is one cluster about a pathology, well differentiated from the rest, that are about hay use in farming.

A similar clustering effect occurs with respect to VIRUS in its biological sense contrasted to the malicious code interpretation; PASCAL as person and as programming language; NLP as acronym for *Natural Language Processing* and as acronym for *Neuro-Linguistic Programming*, and so on.

4.3 Term Translation

A quite different application of the proposed technique is to obtain the translation of a given query term to another language. Let us assume that DUCKBILL PLATYPUS was a term not yet available in our bilingual dictionary.

The resulting network of our algorithm for such entry includes frequent words which can be considered as basic vocabulary, e.g., *mammal*: *mamífero*, *swimming animal*:*animal acuático, eggs*:*huevos*). A new search with these translations, this time in the Spanish web, gives rise to *ornitorrinco* as the most significant MI score. Applying the same strategy we found the Spanish equivalent of *West Nile Virus*. Thus, taking first this term as query term in the English web, we obtain easy translation words such as *mosquito*, *horse*, *infection*, and *transmitted.* In a second search that uses the Spanish translations of these terms, the term *virus del Nilo Occidental* emerges. Analogously, with *model reader*, in the context of semiotics, as translation equivalent of the Spanish *lector modelo*, and *receiver* as the equivalent of Sp. *destino* in the context of the communication theory (Figure 3).

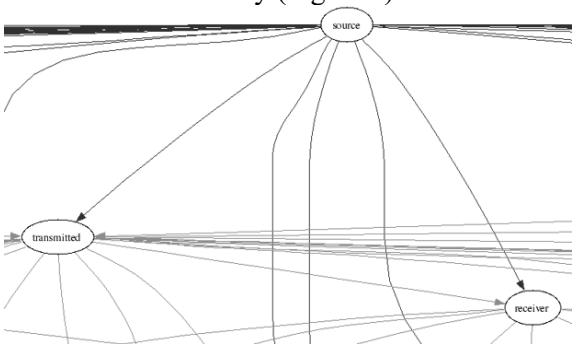

Figure 3: Network for SOURCE to find RECEIVER

4.4 Term Localization

The same strategy applies to localization. Let us assume that a Spaniard wants to know the

equivalent of *aguacate* (*avocado*) in Argentinean Spanish. Searching AGUACATE he/she will obtain the term *persea americana* as one of the most significant collocates. A second search with *persea americana* in combination with the words *nombre* (*name*) and *Argentina* suggests *palta* as the most obvious candidate (we can discard *spp* as a possible translation). Cf. Table 1 for the frequency rank.

Freq. rank	Term
1	*aguacate*
2	*spp*
3	*palta*
4	*nombre*
5	*méxico*
6	*lauraceae*
7	*familia*
8	*argentina*
...	...

Table 1: Collocates of PERSEA AMERICANA NOMBRE - ARGENTINA

(4) is a typical sentence encountered in the retrieved document collection:

(4) *La palta, cuyo nombre científico es persea americana, es de la familia de las Laureáceas, tiene su origen en México, ...*

4.5 Term retrieval

We also tested the algorithm for term retrieval, which addresses the well-known "tip-of-the-tongue" phenomenon: speakers often forget a term but still perfectly recall the purpose of the underlying concept or even the definition of the term in question.

MI rank	Term
1	*acid*
2	*catalytic*
3	*enzyme*
4	*hydrogen*
5	*oxide*
...	...

Table 2: Collocates of CATALYST and GLUCOSE

Let us assume that a speaker searches for the name of the catalyst that helps to break down starch into glucose. Taking CATALYST and GLUCOSE as query terms, the user obtains a network that suggests that *enzyme* is a frequent collocate of both (Table 2).

5 Preliminary Evaluation

From all the envisaged tasks mentioned in the previous section, we are particularly interested in bilingual lexicon extraction, because, in spite of its character, it does not require parallel corpora. Given an entry in a source language, the system returns a ranked list of candidates for translation in a target language.

Thinking of a tool for translators, we do not worry if the correct translation is not the first candidate, because a user, with his or her knowledge, may choose an appropriate translation from a short list. It is easier to recognize a word than to remember it and, even if it is a word the user did not know before, then he or she may observe morphological similarities as a clue in the case of cognates.

We conducted thus a preliminary evaluation, only to estimate overall accuracy, with a multilingual database of names of birds (Scory, 1997). We took a random sample of 25 entries from a total of 700 and entered one by one the names of the birds in English to obtain, with our method, a list of the best candidates for translation in Spanish. The procedure is simple: it takes the best collocate of the query and repeats the search with it in the Spanish corpus. We checked whether the translation provided by the database was among the first three candidates in the list proposed by the system, and depending on it we determined success or failure of the trial.

The study showed 72% coincidence with the database. However, if we consider the non-normative terms as correct (they can be adequate in some contexts), precision raises to 84%. Most often, the failure was due to insufficient data. Some of the species are very rare and it is hard to find documents in Spanish about them. In some of the failed trials the correct candidate was too low in the list returned by the system, or was not present at all. Table 3 shows the results of the experiment. The first and second columns show the English and Spanish names provided by the database, and the third column shows the translation proposed by our method.

Scory's English names:	Scory's Spanish names:	Our method:
firecrest	*reyezuelo listado*	*reyezuelo listado*
brent goose	*barnacla carinegra*	*barnacla de cara negra; ganso de collar*
curlew sandpiper	*correlimos zarapitín*	*correlimos zarapitín*
long-tailed duck	*havelda*	*pato havelda*
short-eared owl	*lechuza campestre*	*lechuza campestre; búho campestre*
song thrush	*zorzal común*	*zorzal común*
pied wagtail	*lavandera de yarrell*	*lavandera blanca*
chaffinch	*pinzón del hierro*	*pinzón vulgar; pinzón común*
stock dove	*paloma zurita*	*paloma zurita*
montagu's harrier	*aguilucho cenizo*	*aguilucho cenizo*
oystercatcher	*ostrero*	*ostrero*
whites thrush	*zorzal*	*zorzal*
short-toed lark	*terrera común*	*terrera común*
kentish plover	*chorlitejo patinegro*	*chorlitejo patinegro*
twite	*pardillo piquigualdo*	*pardillo piquigualdo*
wood pigeon	*paloma torcaz*	*paloma torcaz*
semi-collared flycatcher	*papamoscas semicollarino*	*papamoscas semicollarino*
coot	*focha común*	*focha americana; gallareta americana*
elegant tern	*charrán elegante*	*charrán elegante*
black-necked grebe	*zampuln cuellinegro*	*zampuln cuellinegro*
brown thrasher	*sinsonte castaño*	*sinsonte*
king eider	*eider real*	-
sombre tit	*carbonero lugubre*	-
blyth's pipit	*bisbita de blyth*	-
lanceolated warbler	*buscarla lanceolada*	-

Table 3: Evaluation of the results

Scory's database is incomplete and we were able to find some missing names, as well as other variants from the different variations of the geographically extended Spanish language. For example, *Booted Eagle* can be *águila calzada* or *aguililla calzada*; the *Northern Oriol* should be *Ictérido anaranjado* but the variant *turpial norteño* is also used, the same with *Dark-eyed Junco*, that should be translated as *Cingolo pizarroso*, but in some variants of Spanish it is called *junco ojioscuro. Grey-tailed Tattler* is translated as *Archibebe gris,* but we found *playero de siberia* (in French it is *Chevalier de Sibérie*). This term variation is a problem for the measure of precision, because we are then evaluating not only the performance of the algorithm, but also the difference that exists between normative terminology and real use.

6 Related Work

There are many works that represent the meaning of a term as a network of interdependent nodes labeled by terms, related by edges labeled by predicates. This is the idea behind the Concept Maps (Novak and Cañas, 2006); the Topic Maps (Rath, 1999; Park and Hunting, 2003); the Semantic Web (Shadbolt et al., 2006), among others. Other formalisms, such as semantic networks, may be used to represent concepts and their relationships. A lexical database of such as WordNet (Fellbaum, 1998) is a well known example.

Given the popularity of a search engine such as Kartoo.com (Baleydier and Baleydier, 2006), of the VisualThesaurus.com (Thinkmap Inc., 2004), of a graphical version of Google (Shapiro, 2001) as well as of a variety of other similar representations (Dodge, 2004; Lima, 2005), the idea of a conceptual structure as a net of interdependent nodes is already in the visual imagery of the society. All these representations have in common the goal to transform knowledge serially encoded in text into a topographic structure.

The work related to the automatic generation of conceptual structures involves two fields: term extraction and conceptual relation extraction. For the former, there are several techniques not mentioned in this paper (Vivaldi, 2001, for an overview). For the later, there is also a large body of work.

It is possible to extract semantic relations searching for sentential patterns that provide evidence that between the units X and Y the relation Z holds. For example, X being hyponym of Y, common pattern of this type are *<X>is a type of <Y>*, or *<Y> such as <X>*; *<W>,<X>, and other <Y>,* etc. It is also possible to infer taxonomies from patterns of term variation, for example by the inference that *artificial intelligence* is a kind of

intelligence. Many authors advocate a symbolic approach of this kind; cf., among others, (Hearst, 1992; Godby et al, 1999; Sowa, 2000; Popping, 2000; Ibekwe-SanJuan and SanJuan, 2004).

A different strand uses statistical methods for the extraction of association between terms. Studies of syntagmatic cooccurence for collocation extraction are Church and Hanks (1991); Evert (2004); Kilgarriff et. al (2004); Wanner et al. (2006); among others. Studies of paradigmatic similarity based on vector comparison include Grefenstette (1994); Shütze and Pedersen (1997); Curran (2004). These studies are based on the distributional hypothesis that similar words appear in similar contexts. Studies on graphs drawn by cooccurence data include Phillips (1985); Williams (1998); Magnusson and Vanharanta, (2003); Böhm et al. (2004); Widdows, (2004) and Veronis (2004). Use of graphs is an efficient method in tasks like word disambiguation. By detecting hubs in the graphs, word senses can be determined in a text collection without resort to dictionaries.

7 Conclusions and future work

We have presented a technique for the analysis of concepts and their relations from a purely statistical point of view, without use of direct human judgment or any compiled knowledge from the domain or the language. As a useful metaphor, what we do is to take a picture of the meaning of a term. However, it is also an explicative model as it proposes a reason why it is possible that this technique works, and it is predictive as it has the power to generalize to different contexts and languages.

We contribute to the studies on word cooccurrence in several areas. Contrary to cited authors, our approach is language independent. In addition, we use it for concept map generation and a variety of new applications. We also extend it to experimentation with multilingual corpora.

The work offers prospective engineering applications, but it is also a study of terminology in itself, of the behavior of terms, and not of the terminology of a specific language nor domain. This is, therefore, still in the scope of the interests of linguistics.

Future work will evolve in several directions. Foremost, an extensive evaluation is planned. At the present we are about to evaluate our technique by an algorithm that automatically loops through all the records of the birds database and compares them with the translations provided by our system. This will yield better estimations. We also plan to evaluate the concept maps obtained from the queries with expert users of different areas. Another direction of improvement is a 3D interactive and navigable model of the concept maps since the 2D model entails visualization difficulties. Finally, a web-based version of the prototypical implementation of the technique will be made available soon for free consultation.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. This paper was supported by the ADQUA scholarship granted to the first author by the Government of Catalonia, Spain, according to the resolution UNI/772/2003.

8 References

Baleydier, L and N. Baleydier. 2006. Introducing Kartoo. KARTOO SA. http://www.kartoo.net/e/eng/doc/introducing_kartoo.pdf [accessed April 2007].

Böhm, K., L. Maicher, H. Witschel, A. Carradori. 2004. Moving Topic Maps to Mainstream - Integration of Topic Map Generation in the User's Working Environment. In: J.UCS, Proceedings of I-KNOW'04.241-251

Buitelaar, P., P. Cimiano, B. Magnini. 2005. Ontology Learning from Text: An Overview. In Buitelaar, Cimiano and Magnini (Eds.), Ontology Learning from Text: Methods, Applications and Evaluation,3-12, IOS Press.

Church, K. and P. Hanks. 1991. Word Association Norms, Mutual Information and Lexicography, Computational Linguistics, 16(1):22-29.

Curran, J. (2004). From Distributional to Semantic Similarity. PhD thesis, University of Edinburgh.

Dodge, M. 2007. An Atlas of Cyberspaces: Topology of Maps of Elements of Cyberspace. http://www.cybergeography.org/atlas/topology.html [accessed April 2007].

Eco, U 1981. Lector in fabula la cooperación interpretativa en el texto narrativo, Barcelona, Lumen.

Evert, S. (2004); The Statistics of Word Coocurrences; PhD Thesis; IMS; University of Stuttgart.

Godby, C.; E. Miller, and R. Reighart. 1999. Automatically Generated Topic Maps of World Wide Web Resources. OCLC Library.

Grefenstette, G. (1994) Explorations in Automatic Thesaurus Discovery, Kluwer Academic Publishers, Norwell, MA.

Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. MIT Press.

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics.

Ibekwe-Sanjuan, F. and E. Sanjuan, 2004. Mapping the structure of research topics through term variant clustering: the TermWatch system; JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelles.

Kilgarriff, A. P. Rychly. P. Smrz. D. Tugwell. 2004. The Sketch Engine. Proceedings EURALEX 2004, Lorient, France.

Lima, M. (2005); "Visualcomplexity" [http://www.visualcomplexity.com/vc/ accessed June 2007]

Magnusson, C. and H. Vanharanta. 2003. Visualizing Sequences of Texts Using Collocational Networks. In P. Perner and A . Rosenfeld (Eds).276-283. Springer-Verlag, Berlin, Heidelberg.

Miller, G.A. Virtual meaning. 1995. In *Gothenburg Papers in Theoretical Linguistics* 75:3 – 61.

Novak, J. and A. J. Cañas. 2006. The Theory Underlying Concept Maps and How To Construct Them. Technical Report IHMC CmapTools 2006-01, Florida Institute for Human and Machine Cognition.

Park, J. and S. Hunting. 2003. XML Topic Maps: creating and using topic maps for the Web. Boston, Addison-Wesley cop.

Phillips, M. (1985); Aspects of Text Structure: An Investigation of the Lexical Organization of Text. North-Holland, Amsterdam

Popping, R. 2000. Computer - assisted Text Analysis, London, Sage.

Quasthoff, U., M. Richter, and C. Biemann 2006. Corpus portal for search in monolingual corpora. In: Proceedings of the LREC 2006, Genoa, Italy.

Rath, H. 1999. Technical Issues on Topic Maps, STEP Electronic Publishing Solutions GmbH.

Schütze, H. and J. Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. Information Processing and Management. 33(3):307-318.

Scory, S. 1997. Bird Names, A Translation Index. Management Unit of the North Sea Mathematical Models and the Scheldt estuary, Royal Belgian Institute of Natural Sciences (RBINS). [http://www.mumm.ac.be/~serge/birds/ accessed June 2007]

Shadbolt, N. T. Berners-lee and W. Hall. 2006. The Semantic Web Revisited. IEEE Intelligent Systems 21(3):96-101, May/June

Shapiro, A. 2001. TouchGraph AmazonBrowser V1.01. TouchGraph. http://www.touchgraph.com/TGAmazonBrowser.html (accessed April 2007).

Sowa, J. 2000. Knowledge representation logical, philosophical, and computational foundations, Pacific Grove Brooks/Cole cop.

Thinkmap Inc. 2004. VisualThesaurus.com http://www.visualthesaurus.com (accessed April 2007).

Veronis, J. 2004. HyperLex: Lexical Cartography for Information Retrieval. Computer Speech & Language, 18(3):223-252.

Vivaldi, J. 2001. Extracción de candidatos a término mediante combinación de estrategias heterogéneas. Barcelona: IULA, Universitat Pompeu Fabra, Sèrie Tesis 9.

Wanner, L.; Bohnet, B. and Giereth, M. 2006. Making Sense of Collocations. Computer Speech & Language 20(4):609-624.

Widdows, D. (2004) Geometry and Meaning, Center for the Study of Language and Information/SRI.

Williams, G. 1998. Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. International Journal of Corpus Linguistics 3(1):151-71.