Multi-Word Expressions in the Spanish Bhagavad Gita, Extracted with Local Grammars Based on Semantic Classes

Daniel Stein

Multi-Word Expressions in the Spanish Bhagavad Gita, Extracted with Local Grammars Based on Semantic Classes

Daniel Stein
Hamburg Centre for Language Corpora, University of Hamburg
Max-Brauer-Allee 20, 22765 Hamburg, Germany
ds@daniel-stein.com

Abstract

As religious texts are often composed in metaphorical and lyrical language, the role of multi-word expressions (MWE) can be considered even more important than usually for automatic processing. Therefore a method of extracting MWE is needed, that is capable of dealing with this complexity. Because of its ability to model various linguistic phenomena with the help of syntactical and lexical context, the approach of Local Grammars by Maurice Gross seems promising in this regard. For the study described in this paper I evaluate the use of this method on the basis of a Spanish version of the Hindu poem Bhagavad Gita. The search will be refined on nominal MWE, i.e. nominal compounds and frozen expressions with two or three main elements. Furthermore, the analysis is based on a set of semantic classes for abstract nouns, especially on the semantical class "phenomenon". In this article, the theory and application of Local Grammars is described, and the very first results are discussed in detail.

Keywords: Religious Texts, Multiword-Expression, Local Grammars

1. Introduction

The *Bhagavad Gita*[1] (short: Gita) is one of the central spiritual texts of Hinduism. It is considered to be a synthesis of several religious and philosophical schools of ancient India, including the *Upanishads*, the *Vedas*, and *Yoga*. In Hinduism, a distinction between revealed (*Śruti*) and remembered (*Smriti*) sacred texts is common. But the classification of the Gita depends on the relative branch of Hinduism, as it is a part of a collection of remembered texts as well as a revelation of the god *Krishna*.

In a greater context, the Gita represents the chapters 25-42 of the epic poem Mahabharata ("Great India"). Vyasa, the compiler of the poem comes also into question as the author. The content deals mainly with the conversation between the embodied god Krishna and the prince *Arjuna* on a battlefield inside a chariot. Arunja becomes doubtful when he gets aware that members of his family are fighting for the opposite side. Lord Krishna uses philosophical and spiritual arguments to induce Arjuna to face this war nevertheless. For many commentators, this war stands as an allegory for life itself or the human weaknesses one has to fight, but literal interpretations are also present in the commentaries.

2. Linguistic Analysis

From a linguistic point of view, the Bhagavad Gita is a poem composed of 700 verses in 18 chapters. As this work is a first case study for my dissertation project (that deals with Spanish MWE in various domain texts), I use a Spanish version of the Gita for this analysis. It is published online by the *International Society of Krishna Consciousness* (ISKCON) [2] under the name "El Bhagavad-gita Tal Como Es".

As supposed to be common religious texts or poems, also the Gita is a tight and profoundly complex work with very metaphorical language. Due to the fact that it is a junction of different religious systems, the content of the Gita sometimes seems to be contradictory, e.g. in relation to the question of duality or unitary of being.

Another problem (not only for the analysis) arising for religious texts in general is the questionable possibility of using translations as source: As a matter of fact, a translation always contains an interpretation and (more or less slight) loss of information. Furthermore, a given language's metaphorical character is based on the lexical inventory of this language. So a translation can make it impossible to use the same metaphorical images. This is especially the case for translations of lyrical texts in which not only semantic information needs to be translated, but also measure and rhyme may be essential parts of understanding. In religious texts this may result in serious difficulties. Nevertheless many religious communities accept translated versions of their sacred texts (at least to offer them in a first place to other cultures) and so does Hinduism, at least those branches that offer translations to other languages.[3]

For a non-Hindu reader a further problem is the large number of proper names and even whole semantic concepts with ancient Indian background that are not necessarily translatable to languages with another cultural background (e.g. "*Dharma*" which can roughly be translated as "*that which upholds, supports or maintains the regulatory order of the universe*" [4] and it becomes

[1] In Sanskrit भगवद्गीत "Song of God"

[2] The Spanish version is based on the English translation "The Bhagavad - Gītā As It Is" by Abhay Charan Bhaktivedanta Swami Prabhupada, the founder of the ISKCON. It was first published in 1968 and was the foundation for translations into about 60 languages. Next to the translated text it contains the Sanskrit original, a transliteration and extensive commentary. http://www.spiritual-revolutionary.com/Espanol/BG_espanol/BG_Indice.htm (Last visit: 01/26/2012)

[3] Although this very translation by the ISKCON is one of the widest spread of the Gita, it is still not uncriticized, cf. http://www.dvaita.org/shaastra/gita/prabhupada_review.shtml (Last visit: 26/01/2012).

[4] According to http://en.wikipedia.org/wiki/Dharma (Last visit: 01/24/12).

even more complex if one considers the different meaning in the context of Buddhism).

3. MWE in Religious Texts

In Natural Language Processing, there are several methods and tools that may be useful for the study of religious texts like the Bhagavad Gita and their translations. This includes methods for an automatic alignment of the original source with translated versions, or semantic search engines (e.g. Shazadi, 2011).

But in order to get these methods to work, a text analysis that copes with the complexity of these metaphorical rich and semantically complex texts is necessary. In this specific context, multi-word expressions presumably may play an (even more?) vital role as they already do in nearly every part of NLP. The idiomatic structure and the high level of information density in this text form is presumably very rich in MWE. This information needs to be known for satisfying automatic analyses.

The term MWE regularly subsumes a broad scope of linguistic phenomena [5], so it is useful to choose a well-defined part for the first research.

A categorization of MW can be based on one or more different features, e.g. the word forms, number or syntactical structure of the main elements, the distribution (collocations) etc.

For this paper I'm going to follow the categorization by Guenthner and Blanco Escoda (2004). I'm focussing on nominal MWE which includes the types Compound Nouns (e.g. *agujero negro* = black hole) and Frozen Modifiers (e.g. *miedo cerval* = terrible fear).

A further useful semantical distinction can be made between nouns that describe facts (e.g. *pecado mortal* = deadly sin), and nouns that describe entities (e.g. *obispo auxiliar* = auxiliary bishop).

These categories have been classified in detail for Spanish in the context of the *Nuevo diccionario histórico de la lengua española* (New Historical Dictionary of the Spanish Language) by (Blanco Escoda, 2010). Research has proven semantic categories may vary between languages, so it is recommendable to use a classification that was created for Spanish. I use the semantical class of facts (Spanish: *hechos*). This classification is named *Spanish Hierarchy of Semantic Labels of Facts* (SHSL_F). The MWE that belong into this category are supposed to be easier to identify via context (regarding e.g. the use of auxiliary verbs) than those of the semantical class of entities. The SHSL_F is divided into the following 17 categories:

- Acción (Action)
- Acontecimiento (Event)
- Actitud (Attitude)
- Actividad (Activity)
- Cantidad (Quantity)
- Característica (Characteristics)
- Comportamiento (Behavior)
- Conjunto de Hechos (Set of facts)
- Costumbre (Habit)
- Estado (State)
- Fenómeno (Phenomenon)
- Grado (Degree)
- Parámetro (Parameter)
- Período (Period)
- Proceso (Process)
- Relación Factual (Factual Relationship)
- Situación (Situation)

For the study in this paper I focus on phenomena that may be perceived via the sensual organs or that can be understood. A complete graph of the semantic class of phenomena would include also other aspects e.g. physiological phenomena like a pulse.

4. Tools

4.1 Unitex

The software that is used for this analysis is Unitex 2.1[6] by Sébastien Paumier. Unitex is an open source corpus processing tool that allows the construction of so called Local Grammars (Gross, 1997). This formalism is used to model complex linguistic phenomena using syntactical and lexical context and is often visualized as directional graphs, although technically it is a recursive transition network (cf. figure 1). Local Grammars are useful for many issues in NLP, such as lexical disambiguation, representation of lexical variations in dictionaries or information extraction. They also are very useful for the identification of MWE.

4.2 DELA

The use of Local Grammars relies heavily on a special kind of dictionary, called DELA (Dictionnaires Electroniques du LADL[7]). In DELA, MWE are treated exactly the same way as simple lexical units. The structure of a DELA lemma is as follows:

inflected form , canonical form . syntactical code + semantical code : inflectional code / comment

apples,apple.N+conc:p/this is a comment

With the original installation of Unitex, a basic dictionary of Spanish is included that was designed by the fLexSem group from the Universitat Autónoma de Barcelona (Blanco Escoda, 2001). This basic dictionary contains 638,000 simple words. Considering the inflectional character of Spanish, this is a moderate size which is reflected in the lack of even some common words (see below). Nevertheless the basic dictionary is a good starting point for research and is used for this analysis.

[5] This reflects in the high number of denominations used to describe them, e.g. Collocations, Frozen Expressions, Terms or Compounds, to mention just a few.

[6] http://igm.univ-mlv.fr/~unitex/ (Last visit: 01/29/2012), also for an overview of studies based on Unitex.

[7] Laboratoire d'Automatique Documentaire et Linguistique, the institution in Paris where Maurice Gross developed Local Grammars.

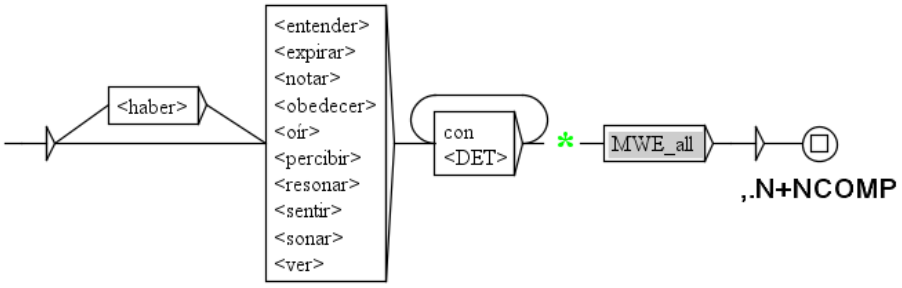

Figure 1: Early version of the Local Grammar for the semantic class "Phenomenon".

Additionally it will be supplemented by some self-constructed dictionaries based on the special features of the Gita.

To make the Gita a useable corpus and to create the mentioned dictionaries, the text needs to be stripped of comments and titles. Also the letter š (which is relevant for the Sanskrit transliterations) has to be added to the Spanish alphabet file in the according Unitex directory. After this step the text has to be loaded as a corpus into Unitex in order to get an overview of its lexical inventory. This can be divided into three lists: The simple and the complex units (that are found in the dictionary as well as in the corpus) and a list of units that just appear in the corpus (and are considered as unknown).

In the case of the Gita, there are 4437 simple-word lexical entries, 0 compounds and 392 unknown forms. 222 items of the unknown forms are regular Spanish words that are not contained in the basic dictionary (e.g. *adelantada* (advanced), *confundida* (confused) but also *yoga*[8]), the other 170 items consist of a special vocabulary of transliterated Sanskrit words which are mostly names (as *Arunja*) with a few exceptions (as the "sacred syllable" *óm*). It is reasonable to create two dictionaries out of this list, one for the Spanish words that are unknown to the basic dictionary to make them useable for further investigations and one just for the Sanskrit words limited for the use with the Gita or similar texts as a corpus.

Corpus processing with Unitex works best with no unknown words so this is an essential issue and easy to accomplish for a short text like the Gita. After applying the newly generated dictionaries on the Corpus in Unitex, it possesses the features presented in table 1.

	Bhagavad Gita
Tokens	53341
Simple word lexical entries	4829
Compound lexical entries	0
Unknown simple words	0

Table 1: Lexical Units in "El Bhagavad-gita Tal Como Es"

[8] Additionally I corrected the misspelled words I found in the Gita in order to improve the analysis, e.g. ahbía -> *había*.

4.3 Local Grammars

The functionality of Local Grammars as a way to extract MWE will be explained using the example graph in figure 1 and the corresponding outputs in figure 2. In simple terms, the Local Grammar in figure 1 matches a phrase, that may begin with an inflected form of the word *haber* (which is part of a perfect tense construction) or with an inflected form of one of the verbs in the list, followed by the word *con* or a determiner or both and at least it needs to find its way through a set of graphs called MWE_all.

Example phrase and recognized path[9]:

He	*entendio*	*el*	*agujero*	*negro*
<haber>	**<entender>**	**<DET>**	**N**	**A**

The graph shown in figure 1 is relatively simple but sufficient to explain the basic principles of Local Grammars and will now be analysed in detail:

- The arrow on the left side represents the starting point; the square within a circle on the right is the end point.
- A text analysis results in a concordance of a number of passages of text that matches to the Local Grammar. A match is achieved when a text passage is found, that is able to be reconstructed by a valid connection from the start point to the end point of the Local Grammar.
- Words without brackets match in cases where exactly the same form is found in the corpus, e.g. *con* (with).
- Words in angle brackets match all inflected forms of this word. E.g. the expression <*entender*> (to understand) matches *entender*, *entiendo*, *entendemos* etc.
- Semantic or syntactic codes in brackets as <DET> represent all words with the same code in the lexicon (here: DET = determiner, e.g. *el*, *la*, *esto*… (he, she, this…).
- Loops are also possible to deal with recursion or some kinds of variation (e.g. this loop matches for *con*, *el*, *con el*, *el con la;* etc:).

[9] N = Noun, A = Adjective

nal se hallan en el mismo nivel, ve las cosas tal como son. La mera renuncia a todas las activi
o es muy inteligente y no puede ver las cosas tal como son. Aquel que no es movido por el ego f
jaya dijo: ¡Oh, Rey!, después de ver el ejército dispuesto en formación militar por los hijos d
o la Superalma. Aquel que entienda esta filosofía relativa a la naturaleza material, la entidad
uerreros Kurus!, nadie había visto esta forma universal Mía antes que tú, ya que ni con el estu
has perturbado y confundido al ver este horrible aspecto Mío. Que ahora se acabe. Devoto Mío, q
lo que es la inacción. Aquel que ve la inacción en la acción, y la acción en la inacción, es i
tud del conocimiento verdadero, ven con la misma visión a un manso y erudito brahmana, a una va
Mi amigo, y puedes por ello entender el misterio trascendental de la misma. Arjuna dijo: Vivasv
arastra: ¡Oh, Rey!, después de oír esas palabras de labios de la Suprema Personalidad de Dios,
os poderosos brazos!, deseo entender el propósito de la renunciación y de la orden de vida de r
querida. La forma que estás viendo con tus ojos trascendentales, no se puede entender simpleme
Sol y la Luna son Tus ojos. Te veo con un fuego ardiente que Te sale de la boca, quemando todo
la comparación con su propio ser, ve la verdadera igualdad de todos los seres tanto en su felic

Figure 2: Concordance for the graph in figure 1.

- The grey shaded box is a link to another graph, in this case to a set of sub graphs that need to be matched, too, in order to achieve a match in the whole graph. The linked grammars are a set that represents all potential possible syntactical variations [10] of MWE in Spanish. The combination of contextual and syntactical Local Grammars is essential for this approach. One can say that the graph in figure 1 is the context that allows the MWE_all main graph to match with higher precision. To get an overview of all the graphs involved in a complete analysis see figure 3.
- It is possible to define left and/or right context that will not be part of the output. The asterisk defines the end of the left context, so everything to the right side of the asterisk is considered as output, i.e. the MWE itself.
- The graph also can be used as a transducer that inserts text into the output. The bold black text beneath the arrow on the right is DELA encoding and can be attached automatically to the output e.g. to create a DELA dictionary of the found compound forms.

As the graph is still in development, this version is considered as a rough sketch and far from the complexity a Local Grammar can reach. After the application of the graph to the "Bhagavad Gita – Tal Como Es" corpus, the 14 passages listed in the concordance in figure 2 are the result. Via application of all the graphs that are not yet involved in the process (cf. figure 3), a significantly larger number of results can be expected.

5. MWE from Bhagavad Gita

The results displayed in figure 2 are of a surprisingly high quality, at least considering the simplicity of the underlying Local Grammar and the complexity of the corpus. When examined in detail, there are still many errors in the expressions found. This implies that the approach of using auxiliary verbs for a Local Grammar to extract MWE of facts is useful. Although a semantic analysis of the phrases obtained would be of interest e.g. in regards to automatic tagging, it is not in the scope of this paper. So I will only analyse whether the expressions in the concordance are MWE and of what syntactical structure they are. I will use a broad definition of MWE including those that are idiomatic as well as those that are common but not obligatory (or fixed) combinations. A helpful rule of thumb is the question if a group of words could possibly be expressed using a single word in another language. An analysis of the phrases reveals the following:

1. ~~*cosas tal como son*~~ (the things as they are): No nominal MWE. The verb *son* is recognized as a

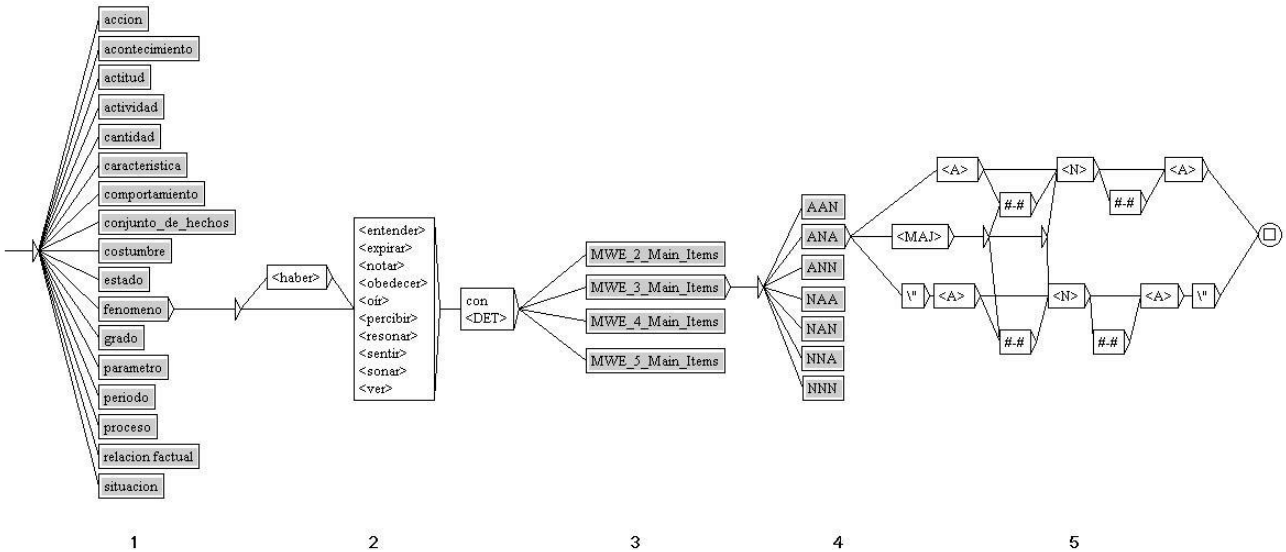

Figure 3: Exemplified combination of the linked Local Grammars, (1) the semantic classes, (2) the class "fenomeno", (3) the MWE overview, (4) MWE with three main items and (5) the MWE with ANA structure.

[10] In fact, for this study I limit the possible MWE for such with two or three main elements.

noun. Both interpretations are present in the dictionary because both may be correct in the right context. But in this case it is unwanted and can be corrected by changing the lexicon entry. As it is impossible to change the standard dictionaries of Unitex, it is necessary to create a lexicon which overrides the standard one. This can be achieved by adding a minus at the end of the dictionary's name (e.g. wronggita-.dic) in order to assign a higher priority to it. For details see the Unitex manual (Paumier, 2010). For this study such a dictionary was prepared containing the four most typical problems: the words *a*, *de*, *la* and *y* encoded as nouns.

2. ~~*cosas tal como son*~~: cf. 1
3. *ejercito dispuesto en formación* [*militar*] (army positioned in military formation): NANA – but because of the exclusion of MWE with four basic elements it is not recognized.
4. *filosofía relativa a la naturaleza* [*material*] (philosophy concerning the material nature): NANN, cf. 3.
5. *Forma universal* ~~*Mía*~~ (universal form of myself): two basic elements, NA – *Mía* is not a noun and needs to be changed in the dictionary, too. However, in this special case Mía is capitalized as it is used by Krishna to refer to himself. The capitalization of pronouns may be a special characteristic of religious texts which has to be taken into account in automatic processing.
6. *horible aspecto* ~~*Mío*~~ (horrible aspect of myself): NA, *Mío* also isn't a noun but capitalized for the self-reference by Krishna, too, cf. 5.
7. ~~*un*~~ *fuego ardiente* (a burning fire): NA, but the indefinite article un needs to be corrected in the Local Grammar.
8. *inacción en la acción* (inaction in action): NN.
9. ~~*la*~~ *misma vision* (same vision): AN, but the defined article la needs to be corrected in the Local Grammar.
10. *misterio transcendental* (transcendental mystery): NA.
11. *palabras de labios* (vain words): NN.
12. *propósito de la renunciación* (intention to renunciation): NN.
13. ~~*tus*~~ *ojos transcendentales* (your transcendental eyes), NA, but the possessive pronoun *tus* needs to be corrected in the Local Grammar.
14. *verdadera igualdad* (true equality): AN.

12 out of the 14 expressions actually contain nominal MWEs. The majority of problems can be corrected by adjusting the dictionary in order to prevent unwanted interpretations of ambiguous words to be applied. Two errors are due to the Local Grammars and can be corrected there. From this point of development work can still be done to improve the recognition of the MWE and to refine the Local Grammars for the context of the Gita. Using the same graphs for other texts should also result in useful data, but dictionaries as well as Local Grammars may need adaption. A more comprehensive evaluation of this approach is beyond the scope of this paper. It would require more advanced graphs for more semantical classes in order to be comparable to other MWE identification approaches as e.g. statistical based ones. Due to the early stage of my study, I was not yet able to realize this.

6. Grammar Construction

The workflow to construct Local Grammars is called bootstrapping and consists of alternating refinement and evaluation of the graphs and the concordances they produce. The development of the grammars for this study is based on two main principles. They guide and as well are guided by the author's intuition in order to improve the Local Grammars.

1) For every semantic class of the SHSL_F there is a graph, all put together in the main Local Grammar. Those graphs contain typical context for the respective class as e.g. auxiliary verb constructions or aspect realizations. Typically the development of these graphs begins with a list of verbal expressions that refer to the context as seen in figure 1. All verbs in the big box describe different ways of perceiving phenomena (*seeing* (*ver*), *hearing* (*oir*), but also *understanding* (*endender*), etc.) As seen above, the output of a Local Grammar based on a semantic class not necessarily belongs to that class (e.g. *ejercito dispuesto en formación militar*). This can be ignored if just the extraction of MWE is the goal. If a semantic tagging is desired later, the graphs may need to be refined on this behalf. A manual control is always necessary.
2) The MWE_all graph contains a set of graphs that are meant to reflect all possible syntactical variations of Spanish nominal MWE (Daille, 1994). The set is subdivided into sets for MWE with two, three, four and five main elements. Each subset contains graphs for as much potential syntactical combinations as possible. The example in figure 3 shows the variations for MWE with ANA structure. The upper path connects the three main elements directly or with hyphens (escaped with two #). The bottom path uses quotes at the beginning and the end as additional constraints. The middle path allows the recognition of MWE that begin with roman numerals as *II Guerra Mundial* (World War II). Every local grammar for semantic classes contains links to the MWE_all or several adequate sub graphs which may be addressed separately.

7. Conclusion

I presented a way to extract MWEs using Local Grammars and semantic classes. The approach seems

promising especially for religious texts and their literary complexity. This study was applied to the Spanish version of the Hindu poem Bhagavad Gita. The first results based on a very simple Local Grammar are encouraging in terms of quality as well as in terms of quantity. The fact that the basis of text is a complex religious poem does not seem to be a problem for the analysis. Quite to the contrary, the analysis revealed some MWE with religious reference (e.g. *misterio transcendental*, *propósito de la renunciación*), which is interesting for different methods of further text analysis (see below).

Because the Local Grammars and semantical classes used in this study are very limited, the results are limited, too. But based on this data, far better results can be expected for a complete analysis, which needs to include Local Grammars for all semantical classes as well as a refinement of the existing graphs and dictionaries.

8. Future Work

There are several interesting questions arising from the study presented here. The questions are related to the approach itself as well as to the data obtained. Which semantic classes will be the most productive when analysing the Gita? Will those classes change if other (religious or profane) texts are analysed? Which elements need to be included into the Local Grammars to improve the identification of MWE? And although the possibility of semantical tagging of the MWE based on the classes seems to be obvious, is it possible to refine the graphs in a way that a manual control is no longer obligatory?

The religious character of the MWE points to the fact, that MWE can be used for automatic text classification. Is it also possible to assign an analysed text to its according religious community?

As the study is based on a translated text, the following question arises: Are MWE in Spanish translations of MWE in Sanskrit? Does a statistical machine translation system that is going to translate the Gita improve by applying a language model that is trained with MWE? Is the semantic tagging that would be achieved simply by adding the semantic class sufficient to create a special search engine? Also the analysis of the distribution of MWE of other semantical classes as entities or abstracts is interesting (in comparison as well as alone).

9. References

Blanco Escoda, X. (2001). Les dictionnaires électroniques de l'espagnol (DELASs et DELACs). In: *Lingvisticae Investigationes*, 23, 2, pages 201-218.

Blanco Escoda, X. (2010). Etiquetas semánticas de HECHOS como género próximo en la definición lexicográfica. In *Calvo, C. et al. (ed.): Lexicografía en el ámbito hispánico*, pages 159-178.

Breidt, E. et al. (1996). Local Grammars for the Description of Multi-Word-Lexemes and their Automatic Recognition in Texts.

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In: *The Balancing Act, Combining Symbolic and Statistical Approaches to Language -- Proceedings of the Workshop*, pages 29-36.

Gross, M. (1997). The Construction of Local Grammars. In*: Finite-State Language Processing*, pages 329-354.

Guenthner, F. and Blanco Escoda, X. (2004). Multi-lexemic expressions: An Overview. In Leclère, C. et al. (ed.) *Lexique, Syntaxe et Lexique-Grammaire,* pages 239-252.

Paumier, S. (2010): Unitex User Manual 2.1, http://igm.univ-mlv.fr/~unitex/UnitexManual2.1.pdf (Last visit: 02/05/2012)

Shazadi, N. et al. (2011). Semantic Network based Semantic Search of Religious Repository. In: *International Journal of Computer Applications*, 36, 9.