

Towards an Image-Term Co-occurrence Model for
Multilingual Terminology Alignment and Cross-
Language Image Indexing

Diego A. Burgos Herrera

TESI DOCTORAL UPF / 2014

DIRECTOR DE LA TESI

Dr. Leo Wanner (Departament de Tecnologies de la Informació i les Comunicacions)

INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA



To my beloved wife,
Zoraida, who is happier
than me that I brought this
thesis to an end.

And to my children,
Carolina and Miguelángel,
who grew up with it.

Acknowledgements

Although risking to unforgivably leave out some names, I would like to express my profound gratitude to people and institutions that made possible taking this dissertation to a good conclusion.

I would like to thank my thesis supervisor, Dr. Leo Wanner, who believed in this project from the very beginning and who supported me, advised me and pushed me even in the hardest moments of this process.

The development of the prototype was possible thanks to: Don Jennings, who programmed the modules and the user interface for BC-Trans; Elena Jaramillo, who developed DORIS, the CBIR application used in our prototype; Antonio Tamayo, who adapted DORIS for the command line so that it interacted with the other components; Bell Manrique for her valuable advice and help with the technical description of the software; Tatiana Vega, Alejandro Arroyave, Pedro Patiño, Gabriel Quiroz, and Antonio Tamayo for testing the paper prototype and providing insightful feedback. Thank you also to Jorge Vivaldi, who gave me valuable support with important data for artifact noun analysis, and to Felipe Zuluaga for giving me a hand with the technical headaches of formatting and word processors.

I am very grateful to the Institut Universitari de Lingüística Aplicada, its great staff, and my professors and colleagues there. I particularly appreciate the unconditional support that I always received from Dr. Teresa Cabré, Dr. Mercè Lorente, Dr. Rosa Estopà, and Dr. Nuria Bel.

I am also thankful to Pedro Patiño and Gabriel Quiroz for reading chunks of my work at different moments of it and for sharing with me their valuable thoughts and advice.

Part of this work was carried out thanks to a grant by the Government of Catalonia according to resolution UNI/772/2003 of the Departament d'Universitats, Recerca i Societat de la Informació.

My eternal gratitude to John Jairo Giraldo and Gabriel Quiroz, who encouraged me to set out on this journey that changed my life. And thanks to God for giving me the strength not to faint during harsh times.

Summary

This thesis addresses the potential that the relation between terms and images in multilingual specialized documentation has for glossary compilation, terminology alignment, and image indexing. It takes advantage of the recurrent use of these two modes of communication (i.e., text and images) in digital documents to build a bimodal co-occurrence model which aims at dynamically compiling glossaries of a wider coverage. The model relies on the developments of content-based image retrieval (CBIR) and text processing techniques. CBIR is used to make two images from different origin match, and text processing supports term recognition, artifact noun classification, and image-term association. The model aligns one image with its denominating term from collateral text, and then aligns this image with another image of the same artifact from a different document, which also enables the alignment of the two equivalent denominating terms. The ultimate goal of the model is to tackle the limitations and drawbacks of current static terminological repositories by generating bimodal, bilingual glossaries that reflect real usage, even when terms and images may originate from noisy corpora.

Resumen

Esta tesis enfoca la relación entre términos e imágenes en documentación especializada y su potencial para compilación de glosarios, alineación de terminología e indexación de imágenes. Asimismo, esta investigación se vale del frecuente uso de estos dos modos de comunicación (i.e., texto e imágenes) en documentos digitales para construir un modelo de concurrencia bimodal que guíe la compilación de glosarios de más cobertura. El modelo se basa en los desarrollos de técnicas de recuperación de imágenes por contenido (CBIR) y de procesamiento de texto. Las técnicas de CBIR se usan aquí para conectar dos imágenes de distinto origen, mientras que el procesamiento de texto sustenta las tareas de reconocimiento de términos, clasificación de nombres de artefacto y asociación término-imagen. El modelo asocia una imagen con el término del texto circundante que la denomina y luego alinea esta imagen con otra imagen del mismo artefacto pero que se origina en otro documento, lo cual permite también la alineación de los dos términos equivalentes que denominan los artefactos de las imágenes. El objetivo principal del modelo es contribuir a compensar el estatismo, las limitaciones y las desventajas de los repositorios terminológicos actuales mediante la generación de glosarios bimodales bilingües que reflejen el uso real de los términos, incluso cuando éstos y sus imágenes se originen en corpus problemáticos.

Table of contents

	Page
1. INTRODUCTION.....	16
1.1. The problem.....	16
1.1.1. Corpus-based dictionary analysis.....	20
1.1.2. Usage-based dictionary analysis.....	25
1.2. Thesis proposal: the BC model.....	27
1.2.1. The Bimodal Co-occurrence hypothesis.....	28
1.2.2. Theoretical Background.....	29
1.2.3. Implementing the BC hypothesis.....	36
1.3. Assumptions and hypothesis.....	37
1.3.1. Assumptions.....	37
1.3.2. Hypothesis.....	38
1.4. Objectives and delimitation of scope.....	39
1.4.1. General Objective.....	39
1.4.2. Specific objectives.....	39
1.4.3. Delimitation of scope.....	39
1.5. Fields of application.....	40
1.6. The Structure of the Thesis.....	41
2. STATE-OF-THE-ART.....	44
2.1. Introduction.....	44
2.2. Image- and text-based information retrieval systems.....	44
2.2.1. Image-term alignment.....	46
2.2.2. Content-based image retrieval.....	50
2.2.3. Cross-language image retrieval.....	54
2.2.4. Domain-specific CLIR.....	57
2.3. MWT and artifact noun recognition.....	59
2.3.1. Multi-word term recognition.....	59
2.3.2. Artifact noun recognition.....	64
3. SEARCH SPACE CHARACTERIZATION.....	69
3.1. Outline.....	70
3.2. Data macrostructure.....	71
3.2.1. General figures.....	73
3.2.2. Category delimitation.....	74
3.2.3. Category filtering and distribution.....	77
3.3. Data microstructure.....	82
3.3.1. URL analysis.....	83
3.3.2. Analysis of BC components.....	85
3.4. Attained goals and drawbacks.....	94
3.4.1. Attained goals.....	94
3.4.2. Search space drawbacks.....	96
4. NOMINAL, ARTIFACT MWT RECOGNITION.....	100
4.1. Introduction.....	100
4.2. Nominal MWT recognition.....	101
4.2.1. Preliminary considerations.....	101
4.2.2. Rule-based MWT recognition.....	105
4.2.3. Seed-based bootstrap MWT recognition.....	106
4.2.4. Predefined-categories-based MWT recognition.....	109

4.2.5.	Evaluation of MWT recognition.....	110
4.2.6.	Results	112
4.2.7.	Discussion	118
4.3.	Artifact MWT recognition	121
4.3.1.	Anchor-based selection.....	121
4.3.2.	Noun classification	121
5.	IMPLEMENTATION OF THE BC-MODEL	132
5.1.	Introduction.....	132
5.2.	System overview	132
5.3.	System Design.....	133
5.3.1.	Behavioral Model.....	134
5.3.2.	Structural/Architectural Model:.....	135
5.4.	User interface	138
5.5.	System evaluation	142
5.5.1.	Evaluation of the CBIR component	142
5.5.2.	Evaluation results.....	144
5.6.	Remarks	152
6.	CONCLUSIONS.....	155
6.1.	Hypothesis validation	155
6.2.	Objectives attainment	158
6.3.	Contributions	161
6.3.1.	Characterization of the problem	161
6.3.2.	The BC model	162
6.3.3.	The Search Space	163
6.3.4.	MWT and artifact noun recognition	164
6.3.5.	Prototype	167
6.4.	Limitations	168
6.4.1.	The BC model	168
6.4.2.	MWT and artifact noun recognition	169
6.4.3.	Prototype	169
6.5.	Lines of future research.....	169
	REFERENCES.....	171
	ANNEXES	180

Index of tables

	Page
Table 1. Examples of term variants denoting the same concept.....	18
Table 2. Number of English and Spanish entries analyzed for each dictionary	24
Table 3. In each cell, the number of queries made for two or more words is on the left. The total number of queries including single-word expressions is on the right.	25
Table 4. Best MAP scores for some ImageCLEF tasks - Different modalities.....	56
Table 5. Best MAP scores for some ImageCLEF tasks - Visual modality	56
Table 6. Fields of ODP structure file	72
Table 7. Fields of ODP content file.....	72
Table 8. Most frequent strings for URLs in English and Spanish search space.....	85
Table 9. Main figures of text in BODY section.....	92
Table 10. Main figures of images in the search space. Counting is presented after removing duplicates.....	94
Table 11. ANOVA of Evaluation score - Spanish.....	115
Table 12. ANOVA of Extracted candidates - Spanish	115
Table 13. ANOVA of Evaluation score - English.....	116
Table 14. Precision and recall for English MWT recognition.....	118
Table 15. Precision and recall for Spanish MWT recognition.	118
Table 16. Some examples of the relative edit distance measure.	123
Table 17. Example of lexical typology.....	125
Table 18. Evaluation results.....	127
Table 19. Nouns annotated with WordNet supersenses.	128
Table 20. Results of the semantic annotation.	128
Table 21. Precision and Recall for three lexical sense-based methods of artifact noun annotation, including the baseline.	129
Table 22. MAP scores for the head noun matching task.....	147
Table 23. Best MAP scores for some ImageCLEF tasks - Different modalities.	148
Table 24. Best MAP scores for some ImageCLEF tasks - Visual modality.....	148
Table 25. ANOVA results for feature coordinates mean comparison.....	150
Table 26. Spanish monolingual retrieval.	156
Table 27. Spanish monolingual retrieval 2.....	157
Table 28. Cross-Language Spanish-English retrieval.....	157

Index of figures

	Page
Figure 1. Frequency of documented MWTs	22
Figure 2. Overall documented MWTs.....	22
Figure 3. Documented nominal items vs. total of nominal items according to the number of modifiers (PREM=Premodifiers, HN=Head Noun, N=Noun)	23
Figure 4. Average of tokens per term in the dictionaries.....	24
Figure 5. Mean of retrieved documents for English and Spanish.....	26
Figure 6. Two- to four-token English terms retrieving fewer than 10 documents.....	27
Figure 7. Index term refers to a word or set or words which designate an object in an image.	28
Figure 8. Real world instance of the BC-hypothesis in a bilingual setting.....	29
Figure 9. Image as a universal, language independent representation	32
Figure 10. BC-based methodology for term alignment and image indexing.....	37
Figure 11. An Illustration of image translation, rotation, scaling and deformation. The last frame shows a combination of all of them.....	51
Figure 12. Lemma frequency analysis of keywords and descriptions for English	76
Figure 13. Lemma frequency analysis of keywords and descriptions for Italian.....	76
Figure 14. Distribution of English search space.....	78
Figure 15. Distribution of Spanish search space	80
Figure 16. Instance of the BC Hypothesis in a web catalog.....	83
Figure 17. Most frequent strings in URLs for English search space.	84
Figure 18. Most frequent strings in URLs for Spanish search space.....	84
Figure 19. Lemma distribution of keywords in English search space.	86
Figure 20. Lemma distribution of descriptions in English search space.....	87
Figure 21. Lemma distribution of keywords in Spanish search space.....	87
Figure 22. Lemma distribution of descriptions in Spanish search space.	88
Figure 23. Analysis of prepositions in dictionaries.....	104
Figure 24. Recognized (blue) vs. missed (red) referent MWTs for the three methods in English (left side) and Spanish (right side).....	112
Figure 25. Edit distance mean (red) and Standard deviation σ (blue) for the three methods in English (left side) and Spanish (right side).....	113
Figure 26. Evaluation score mean (red) and Standard deviation σ (blue) for the three methods in English (left side) and Spanish (right side).....	113
Figure 27. Candidates per context. Mean (red) and Standard deviation σ (blue) for the three methods in English (left side) and Spanish (right side).	114
Figure 28. Candidates per context – Spanish	115
Figure 29. Medians and Variance for Spanish Recognition.....	115
Figure 30. Medians and Variance for English Recognition.....	116
Figure 31. Candidates per Context - English	116
Figure 32. Precision and recall for English MWT recognition.	117
Figure 33. Precision and recall for Spanish MWT recognition.....	117
Figure 34. Number of term candidates recognized by each method in in English (left side of the x-axis) and Spanish (right side of the x-axis).	119
Figure 35. Precision and recall for artifact noun annotation with SuperSense Tagger (SST), UKB algorithm, and the most frequent sense (MFS).....	129
Figure 36. Activity diagram.....	135
Figure 37. Component model.....	136

Figure 38. Class diagram.....	137
Figure 39. Deployment Nodes.....	137
Figure 40. User is prompted to enter query.....	138
Figure 41. System returns related terms.	138
Figure 42. System returns candidate translations.....	138
Figure 43. User is prompted to upload an image.	140
Figure 44. System returns term record.....	140
Figure 45. User selects an image to upload.....	140
Figure 46. Select candidate target terms.	141
Figure 47. Term record.....	141
Figure 48. Ranking and distance for image sets A, B and C in the image name matching task.	144
Figure 49. Evaluation results of the FTM task for the three image sets.	145
Figure 50. Evaluation results of the HNM task for the three image sets.	146
Figure 51. Progression of results by task for each image set (A, B, and C).....	147
Figure 52. Feature comparison among sets.....	150
Figure 53. Evaluation through queries.	152
Figure 54. Two artifacts with different textures.....	152
Figure 55. Task 1. Instruction page.....	182
Figure 56. Task 1. Screen 1.....	182
Figure 57. Task 1. Screen 2.....	182
Figure 58. Task 1. Screen 3.....	183
Figure 59. Task 1. Screen 4.....	183
Figure 60. Task 2. Screen 1.....	184
Figure 61. Task 2. Screen 2.....	184
Figure 62. Task 2. Screen 3.....	184
Figure 63. Task 2. Screen 4.....	184
Figure 64. Task 2. Screen 5.....	185
Figure 65. Task 2. Screen 6.....	185
Figure 66. Task 2. Screen 7.....	185
Figure 67. Task 2. Screen 8.....	185
Figure 68. Task 2. Screen 9.....	186

Index of annexes

	Page
Annex 1. English and Spanish syntactic patterns	180
Annex 2. Paper prototype	182

1. INTRODUCTION

This thesis addresses the potential that the relation between terms and images in multilingual specialized documentation has for glossary compilation, terminology alignment, and image indexing. It takes advantage of the recurrent use of these two modes of communication (i.e., text and images) in digital documents to build a bimodal co-occurrence model which aims at dynamically building glossaries of a wider coverage. The model relies on the developments of content-based image retrieval (CBIR) and text processing techniques. CBIR is used to make two images of different origin match, and text processing supports term recognition, artifact noun classification, and image-term association. The model aligns one image with its denominating term from collateral text, and then aligns this image with another image of the same artifact from a different document, which also enables the alignment of the two denominating terms. The ultimate goal of the model is to tackle the limitations and drawbacks of current static terminological repositories by generating bimodal, bilingual glossaries that reflect real usage, even when terms and images may originate from noisy corpora.

The remainder of this chapter elaborates more on the problem that motivates the present research. In order to exemplify the problem, some specialized dictionaries are quantitatively and qualitatively analyzed as for their relevance for the translation of a set of terms extracted from a corpus, and for their usage. Then, the bimodal co-occurrence model, which constitutes the core of this thesis' proposal, is described and its theoretical background presented. Next, the thesis' hypothesis, objectives, limitations, and fields of application are outlined. The chapter ends with a description of the structure of the whole dissertation.

1.1. The problem

Nowadays, a great deal of specialized translation and terminology-based tasks must be carried out on the basis of rather static low-coverage textual terminological resources, e.g., specialized dictionaries, terminological databases, etc. Given the non-dynamic nature of most resources, some (or many) terms tend to become outdated and therefore do not reflect the conventional usage of terms among experts. Likewise, dictionaries and even terminological databases lag behind online and hardcopy product catalogues and technical manuals with regard to new

terminology. As advances in science and new technologies take place, concepts change or emerge and terms show an evolution that is revealed too late in terminological resources.

Because of the above-mentioned low coverage and static conditions, dictionaries lack of the necessary information which directly and indirectly affects the writing and translation of technical and scientific documentation. In the worst-case scenario, translators are not able to locate a suitable translation for a given term, in which case they come up with a new equivalent. However, as the search space for term translations is so big, it is probable that an equivalent already existed. If it did, the simultaneous and extensive practice of proposing new terms will likely result in an undetermined number of terminological variants throughout corpora and databases and it would have consequences in specialized knowledge transference and discussion. The following case illustrates such an effect. Consider the terms below taken from the online Termium Plus®¹:

English terms	French terms
[1,1'-biphenyl]-2-ol	[1,1'-biphényl]-2-ol
biphenyl-2-ol	biphényl-2-ol
Biphenylol	biphénylol
1,1'-biphenyl-2-ol	1,1'-biphényl-2-ol
(1,1-biphenyl)-2-ol	(1,1-biphényl)-2-ol
(1,1'-biphenyl)-2-ol	(1,1'-biphényl)-2-ol
2-biphenylol	2-biphénylol
o-biphenylol	o-biphénylol
2-diphenylol	2-diphénylol
o-diphenylol	o-diphénylol
hydroxybiphenyl	hydroxybiphényle
2-hydroxybiphenyl	2-hydroxybiphényle
o-hydroxybiphenyl	o-hydroxybiphényle
2-hydroxydiphenyl	2-hydroxydiphényle
o-hydroxydiphenyl	o-hydroxydiphényle
ortho-hydroxydiphenylVI	ortho-hydroxydiphényleM,VI
hydroxy-2-phenylbenzene	hydroxy-2-phénylbenzène
phenylphenol	phénylphénol
2-phenylphenol	2-phénylphénol
o-phenylphenol	o-phénylphénol
OPP	orthophénylphénolM,VI
orthophenylphenolVI	xénol

¹ <http://www.termiumplus.gc.ca/>

Xenol	o-xénol
o-xenol	orthoxénol
Orthoxenol	o-xonal
o-xonall	

Table 1. Examples of term variants denoting the same concept

All the above terms represent a single concept: *a white crystalline powder, with light phenol odor, soluble in ethanol, alkalis, greases and oils, not very soluble in water, used like protective agent for fruits and vegetables as well as for the preparation of disinfectant ointments*. In an ideal setting, a concept should be represented by just one term, as stated by the Wüster's General Theory of Terminology, (cf. Cabré, 2003). However, as Cabré (2000) observes, this is not always the case. Terminology variation exists and it implies the possible availability of a number of terms that independently represent one concept, as in the example above.

Freixa (2002, p. 123) examines some issues related to the use of synonyms, that is, terminological variants. She explicitly mentions, among others, a translation-related issue that supports the above-described problem and which motivates the present research:

“The denomination landscape becomes more dispersed when translators of scientific texts propose different translations of the same term.”[our translation] (Freixa, 2002, p. 123)

Freixa (2002) remarks the fact that the same term can be rendered in a different way by each translator, which results in an undetermined number of synonyms. She sees translators not as mere users of terms but also as creators of new variants through the translation process. This could be seen as a cause of term variation since, if a translator is not able to find the right translation equivalent, he/she will look for an alternative translation of the term.

Along the same lines, Gamero (2001, p.43) observes:

When the term search is unsuccessful due to a partial or nonexistent equivalence between languages, a translator can solve the problem via three techniques: loan, neologism and paraphrase. When the decision is made on loan or neologism, the translator should ask a terminology expert for advice since the uncontrolled appearance of loans and neologisms implies a risk for international harmonization of terms and favors the over-generation of synonyms. [our translation]

Various classifications have been proposed to cluster terminological variants according to their morphological, syntactic or semantic variation (Freixa, 2002, p.267-358; Savary and Jacquemin,

2003). However, such classifications just account for the observable terminological *universe*. Only known objects, events or phenomena whose properties are recognized can be classified. Gamero (2001) refers to an unsuccessful search in the observable universe. But, perhaps another onomasiology- or translation-based equivalent lays on a non-observable resource which could be surprising because of its formal configuration and ideal due to the nature of its conception.

Therefore, although terminological variation responds to the nature of language, it is also advisable to control the phenomenon to a reasonable extent. That is, even though variation cannot be completely avoided when writing or translating, efforts should be put into using existing variants –provided that they are widely accepted– instead of creating new ones. However, checking the existence of a term is not a trivial task. Terminologists, translators and technical writers usually resort to dictionaries or related documentation in order to look for denominations, definitions or target language equivalents of terms. But if the search for, say, a target language term is unsuccessful with a reasonable investment of time and resources, then a translation is proposed; a translation which could turn into a variant if another rendering already existed. As an example of the process of translation proposals, let us cite the UPF_Term² which hosts a project called Termium³ with ≈ 675 terminological records and $\approx 3,292$ entries in Catalan, English, French, and Spanish. 362 of the entries in Catalan and Spanish are proposals by the translator or by a consulted specialist.

Some questions arise then concerning such productive processes of translation and terminology: was the search exhaustive enough to assure the term did not exist yet? Are we using the appropriate tools to explore the whole search space, or do such tools and procedures let us see just a part of it?

The causes of the over-generation of variants and of the difficulty to find appropriate term translations must be identified in order to be able to propose a solution to the problem. It could be said, then, that the indirect effect of the problem is the generation of an undetermined number of terminological variants throughout corpora and databases with probable consequences in communication among experts. According to this, the causes can be

² Terminological bank of the Pompeu Fabra University (<http://www.iula.upf.edu/rec/upfterm/cat/index.htm>)

³ As a result of the cooperation with The Government of Canada's terminology and linguistic data bank.

established, on the one hand, as a lack of theoretical models for effective term location and retrieval from unstructured or semi-structured repositories (i.e., the Web); and, on the other hand, as the absence of a practical implementation of strategies to dynamically compile and update wide-coverage terminological dictionaries to support specialized translation and technical documentation development.

The following section presents more detailed empirical evidence related to (1) the capacity of terminological resources to include terms found in corpora; and (2) the frequency of usage of terms documented in dictionaries: a web-based analysis.

1.1.1. Corpus-based dictionary analysis

The analysis below serves to determine the extent to which terms used by experts in specialized texts are documented in dictionaries. Given the requirements of the model presented in §1.2, a domain with a representative number of object referents (e.g., spare parts) was necessary to maximize applicability of the model. That is why texts of automotive engineering were selected for this study.

1.1.1.1. Corpus and data extraction

As a corpus for manual term extraction, six articles were selected from the Tech Briefs section of an issue of *Automotive Engineering International Online*.⁴ The corpus selection was carried out on the basis of specific criteria of specialized discourse and topicality. Given that terms are often nominal multiword terms (MWTs), and that their degree of specialization is determined by the number of modifiers of the head noun, the language unit for this study will be MWTs⁵. Accordingly, 152 MWTs were manually extracted from the corpus, provided that they complied with two basic criteria: 1) they had to be nominal units, in the sense defined, for instance, by Bosque (1999, p. 5-8), and p. 2) they had to denote a countable concrete entity⁶. A sample of the MWTs extracted for the experiment is shown below:

aluminium cylinder head
diesel particulate filter
forged steel crankshaft

⁴ This specialized online journal can be accessed at <http://articles.sae.org/automotive/browse/>.

⁵ See §2.1.2 for a finer definition of MWT.

⁶ See Quirk *et al.* (1985: 247) or Bosque (1999: 8-28, 45-51) for a finer definition of concrete nouns.

front brake rotor
fuel injection nozzle
limited-slip differential
electronic control unit
hydraulic torque converter

The determination of the terminological character of these MWTs might raise questions concerning terminological validation criteria, such as expert opinion, terminographical documentation, and usage frequency, verification in specialized dictionaries and in the web as well as statistical analyses. As a point of departure, a series of tests proposed by Cabré (1993, p. 304-305) were applied to verify that the selected phrasal terms were not merely simple combinations of words but MWTs.

1.1.1.2. Determining dictionary trends

The statistical analysis presented below determines to what extent the MWTs of the sample were documented in three specialized dictionaries and one general technical dictionary. The four dictionaries consulted for this study were: *Diccionario del motor* (Orueta Colorado, 2004), *Diccionario de la automoción* (South & Dwiggins, 1999), *Dictionary of Automotive Engineering* (De Coster, 2003) and *Routledge English Technical Dictionary* (1998). As our objective was to observe generalized trends in terminography rather than evaluating individual dictionaries, the dictionaries are mentioned here but will not be specifically linked to the reported results about specific data contained in each one.

1.1.1.2.1. Dictionary coverage of MWTs

Figure 1 shows the frequency of documented MWTs, i.e. the number of MWTs found in the dictionaries exactly as they were extracted from the corpus. While the percentages vary among dictionaries, the cumulative percentage of documented MWTs in the four dictionaries is 48.7%. This means that 51.3% of the MWTs would not be found as required by translators in these dictionaries.

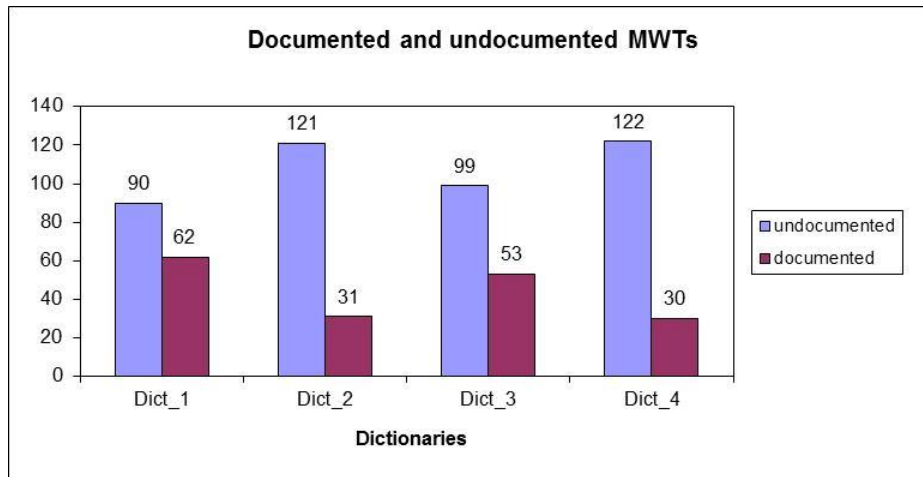


Figure 1. Frequency of documented MWTs

Out of the 152 MWTs, only 14 appeared in all four dictionaries; 21 MWTs appeared in only one of the dictionaries; and the remaining 39 MWTs were found in two or three of the dictionaries, for a total of 74 documented MWTs (see Figure 2). The distribution of the data in the four dictionaries was analyzed and compared to see if there were significant differences. The analysis showed that two clusters resulted, namely, the first one in dictionaries 4 and 2 documenting fewer MWTs, and the second one in dictionaries 1 and 3 which include more MWTs. As expected, this confirmed that some dictionaries offer more or better solutions depending not only on their quality, but also on the degree of specialization. The fact that the source text deals with automotive engineering does not mean that terms belonging to other domains will not cause translation problems. Therefore, a general technical dictionary might provide equivalents for a number of less specific terms in a wide range of subject matters, while a specialized dictionary on automotive engineering offers fewer but more specific terms in its domain.

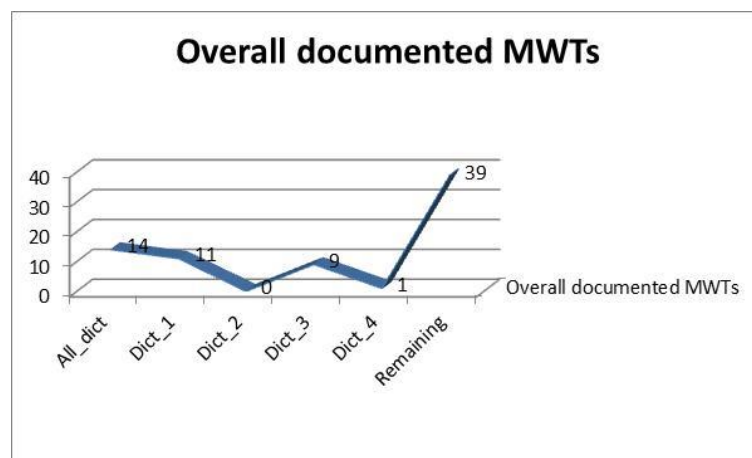


Figure 2. Overall documented MWTs

1.1.1.2.2. Term length in context and in dictionaries

It appears that the number of tokens in a term is decisive for the probability of finding the term in a terminological source. Figure 3 shows a comparison of documented and undocumented MWTs in dictionaries according to the number of modifiers. As illustrated, single-word expressions account for the greatest relative number of the documented nominal lexical items (31 documented nouns out of a total of 32 extracted single nouns). The relative value decreases when modification occurs and the number of modifiers increases. These initial observations also suggest that many of the MWTs extracted from our corpus are two-word expressions and that the second relative value of the documented MWTs is represented by such units (31/56). However, the number of MWTs consisting of three or more tokens is significant even for a small sample size such as the one used in this analysis. If predictions were made on the basis of the fact that most of the terms documented in the dictionaries or terminological databases are two-word expressions, there would be no equivalents for at least 64 of our MWTs in such terminological sources.

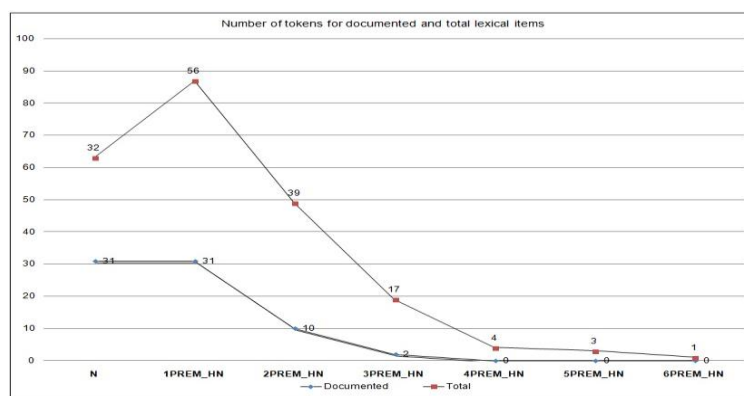


Figure 3. Documented nominal items vs. total of nominal items according to the number of modifiers (PREM=Premodifiers, HN=Head Noun, N=Noun)

1.1.1.2.3. Verifying dictionary trends

In order to confirm our initial observations above with regard to term length in dictionaries, three dictionaries were statistically analyzed to verify the mean of tokens of the terms recorded. As a means of establishing wider generalizations, three dictionaries from different domains were consulted: a) *International Electrotechnical Vocabulary (IEV) online database*, b) *Diccionario de la Automoción (DA)* (South & Dwiggins, 1999), and c) *Spanish Dictionary of Business, Commerce and Finance (BCF)* (1998). See details in Table 2.

<i>Dictionary</i>	<i>English entries</i>	<i>Spanish entries</i>
IEV	15222	16185
DA	3822	3962
BCF	23040	39045
<i>Total</i>	<i>42084</i>	<i>59192</i>

Table 2. Number of English and Spanish entries analyzed for each dictionary

As can be seen in Figure 4, the mean of tokens of English terms in IEV is 2.24, 1.92 in DA and 2.04 in BCF. The cumulative mean of English tokens in all three dictionaries is 2.11. The mean of tokens of Spanish terms in IEV is 3.14, 2.70 in DA and 2.76 in BCF. The cumulative mean of Spanish tokens for all three dictionaries is 2.86. An analysis of the results reveals a significant difference between the means of the number of English and Spanish tokens. If we consider that the analyzed English and Spanish entries in these dictionaries are equivalents of each other, the fact that the average for the Spanish terms is three tokens compared to two tokens for the English terms might be explained by the tendency of translations from English into Spanish to include a preposition, which is often necessary.

These first results show that dictionaries still tend to be very conservative as regards the number of tokens included in MWT's. Such conclusion should not be underestimated for translation. It means that a high number of terms consisting of three or more tokens will have to be found by other means that might not assure terminology consistency or expert approval.

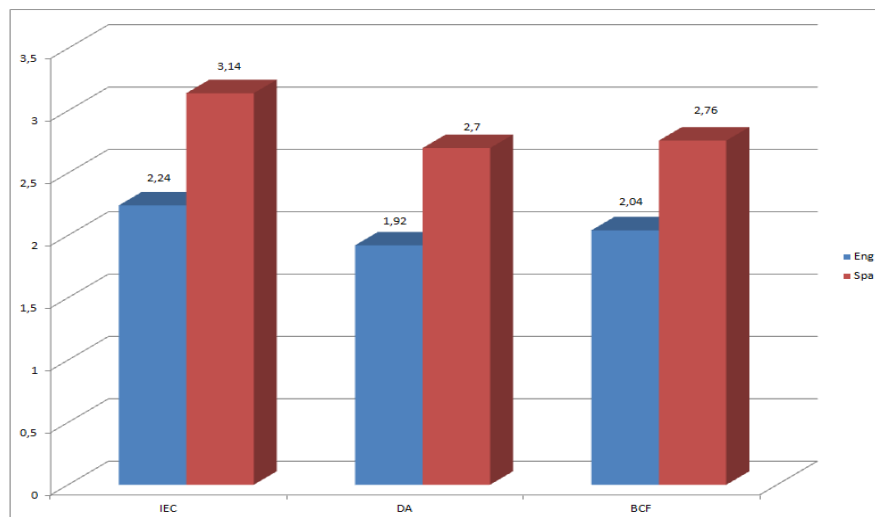


Figure 4. Average of tokens per term in the dictionaries

1.1.2. Usage-based dictionary analysis

A second issue in terminology management in dictionaries concerns the number of synonyms offered as equivalents for a source language term. The following important questions arise: If no distinction is made between the synonymous equivalents (geographical variant, for instance), as is often the case, which one should be used? How have all those synonyms been selected? What criteria have been followed? What is the real usage of such synonyms by the specialist community? The last question is perhaps the most critical one.

Given the practical impossibility of consulting specialists to verify the usage of all the equivalents provided by the dictionaries, another statistical analysis was carried out. Since the Internet is nowadays the world's largest source of information, and therefore reflects to a certain degree the state and evolution of social and scientific knowledge, web queries were made to extract figures about the usage of terms contained in the specialized dictionaries and how this usage is related to the number of tokens per term. For this part of the experiment, Google was queried for as many of the English and Spanish entries from the above-mentioned dictionaries as was technically possible, and the results were statistically analyzed (see details in Table 3). The number of queries guarantees that the addition of new websites or the removal of other ones will not be statistically significant, although the figures could certainly change as the time passes by.

<i>Dictionary</i>	<i>Spanish queries</i>	<i>English queries</i>
IEV	14023/16183	12547/15220
DA	3068/3962	2666/3822
BCF	9432/11921	4701/6727
<i>Total</i>	<i>27277/32066</i>	<i>19914/25769</i>

Table 3. In each cell, the number of queries made for two or more words is on the left. The total number of queries including single-word expressions is on the right.

Even though Internet statistics report a much higher number of English than Spanish users – thus suggesting many more English web pages⁷ –, the difference in the mean of retrieved documents for each language is even more dramatic than should be expected, as shown by Figure 5. To be coherent with the results of the previous analysis, only the documents retrieved with two or more word queries for both languages were analyzed. This strategy also helped to

⁷ Internet statistics do not provide the exact number of pages in a given language (for more information, visit <http://www.internetworldstats.com/>).

avoid outliers produced by single-word expressions which are often used in general language too.

Figure 5 shows that for each query of a lemma from our dictionaries containing two or more words, an average of 7,860 documents were retrieved from the Web for Spanish, and 246,575 documents for English. It is uncertain whether such figures are due only to the number of pages in each language on the web. In this specific case, the ratio is 5.1 English-speaking users per each Spanish-speaking user,⁸ while the ratio of retrieved documents is 31 English documents per each Spanish document.

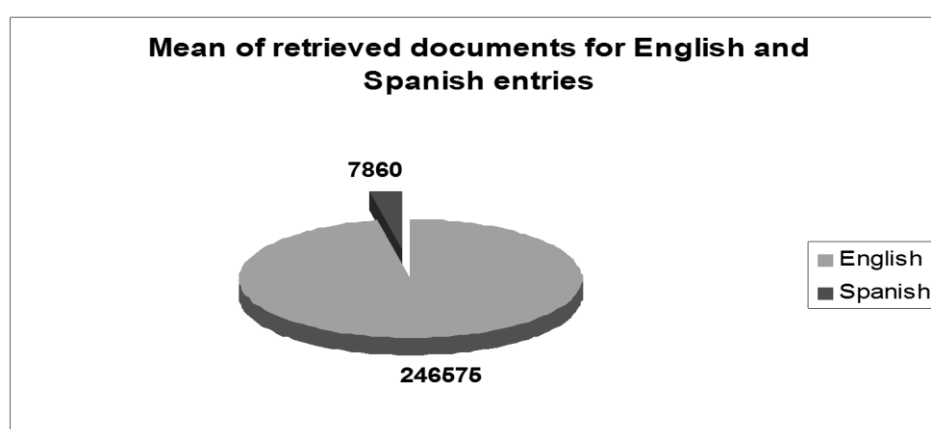


Figure 5. Mean of retrieved documents for English and Spanish

The low number of documents retrieved in Spanish does not necessarily mean that more relevant documents do not exist, but perhaps that the query patterns were not appropriate enough to match the search space patterns. The results of the previous analysis could be interpreted as if most of the Spanish terms documented in the analyzed dictionaries were proposed by a translator or a specialist and not derived from, say, a corpus. The extremely greater number of English documents is therefore not surprising, considering that the English terms were generated first than the Spanish ones.

On the basis of this interpretation, a final analysis was done for the terms in our dictionary sample containing between two and four tokens by means of which fewer than 10 documents

⁸ Users by language in 2005. Figures updated to 2010 show a ratio of 3,5 English-speaking users per each Spanish-speaking user. From 2000 to 2010, growth of English in Internet was of 281,2 % while growth of Spanish in the same period was of 743.2 % (source: Internet World Stats - www.internetworldstats.com/stats7.htm).

were retrieved. An interesting fact regarding English is shown in Figure 6. In the group of terms retrieving fewer than 10 documents, there are many more 3-token terms than 4-token terms. It is certainly intriguing to note that the degree of specialization of the terms represented in the number of tokens correlates positively with the number of retrieved documents.

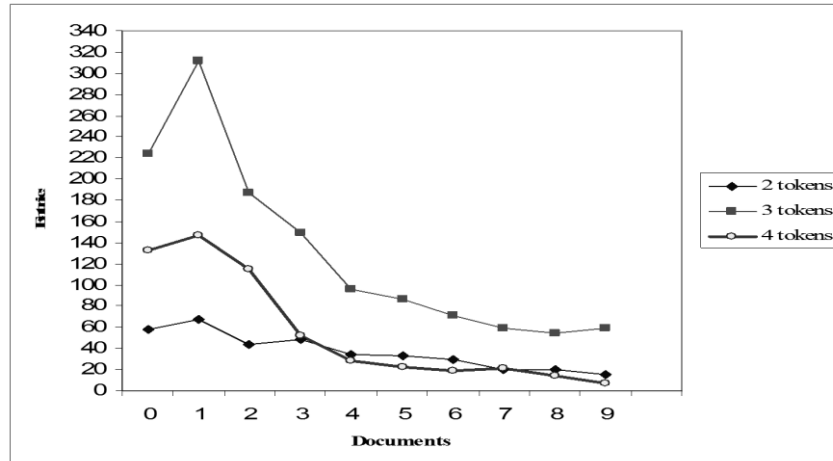


Figure 6. Two- to four-token English terms retrieving fewer than 10 documents

As regards the results for Spanish obtained in the same analysis, there was a total of 7,655 terms retrieving fewer than 10 documents. This raises the question about the reliability of a term found in only 10 or less web documents. Perhaps the authority of such documents could help resolve the issue of reliability. Some could argue that such results could be caused by the fact of querying the web engine with 4-token terms, but what happens with the 341 2-token terms retrieving no documents, or with the 395 terms retrieving just one document? Are they terminological neologisms? Even 3-token terms should retrieve a representative number of documents since, as said before, one of the tokens would probably be a preposition.

1.2. Thesis proposal: the BC model

It is the limitations and drawbacks of current terminological repositories described above that motivate this thesis. As a contribution to tackle the problem, this section presents the core of our proposal, that is, a theoretical construct that we call the bimodal co-occurrence (BC) model, which in turn bases on the bimodal co-occurrence (BC) hypothesis. The subsections below define the BC hypothesis and sketch the thesis' methodology to dynamically build glossaries of a wider coverage based on the BC model.

1.2.1. The Bimodal Co-occurrence hypothesis

We assume language independent bimodal co-occurrence of images and their index terms in the corpus. This implies that if the image of an object occurs in a document of the corpus, the corresponding index term of the object in the image will also occur in the same document (see Figure 7).

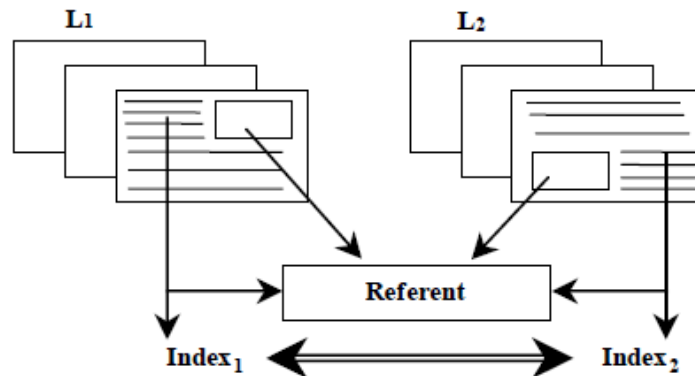


Figure 7. Index term refers to a word or set of words which designate an object in an image.

Figure 7 also suggests the BC in a bilingual setting. That is, when there is an image of an artifact in the source language corpus along with its index term, there should also be an image of the same artifact along with its index term in the target language corpus. This means that the fact of making both images match would get the two equivalent terms closer. As an example, Figure 8 shows screenshots of two documents presenting the BC in a bilingual setting with reference to the same real world artifact.

By visualizing the image and comparing additional information such as brands, measurements, references and features, a user can determine whether the images represent the same artifact. In the case of Figure 8, for example, the user's knowledge and the document layout should be enough to establish that *Regulator* is the English translation of the Spanish term *Regulador de volt.* In such a case, the maximum error of stating that the terms designating the images in each document are translations of each other could not be greater than the error of acknowledging the equivalence of two translations found in a specialized bilingual dictionary.

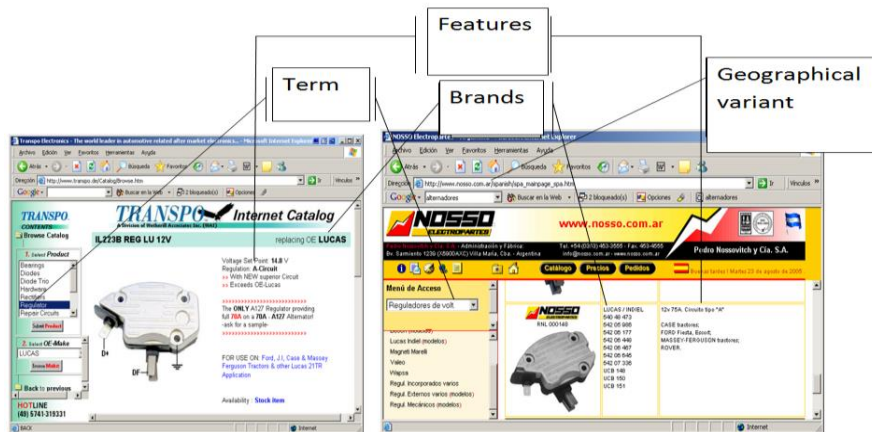


Figure 8. Real world instance of the BC-hypothesis in a bilingual setting.

Similarly, information on geographical variants could also be obtained and considered via domain extensions. For example, in Google, the command `site:.co` could be used to retrieve websites from Colombia and, therefore, to detect the Colombian variant of a term.

The examples above are part of some more representative instances of this BC assumption observed in the frame of this work. The observations were carried out in a preliminary empirical study for English. A sub-corpus of 20 MWTs designating artifacts from the automotive industry was extracted from an issue of the *Automotive Engineering International Online journal's* Tech Briefs section and used to retrieve documents from the web. The first 10 results (i.e., web pages) for each term were stored. Each of the web pages was manually analyzed to check the BC. The 20 terms observed the BC-hypothesis in 145 sites (out of 200) which means 72,5% of positive cases (Burgos and Wanner, 2006).

In this thesis, we set out to (i) prove the validity of the BC hypothesis, and (ii) take advantage of the model derived from the hypothesis to search for translation equivalents.

1.2.2. Theoretical Background

It is clear from the previous description that visual and linguistic expressions of an object are the key components of the model. The sections below explain the theoretical basis as well as some empirical observations that support the choice of these components and describe how both components are related.

1.2.2.1. The visual representation component

Images, as representation of real world entities, constitute a *sine qua non* prerequisite for a number of language-related tasks. For instance, children as well as foreign language learners often resort to images in order to concretize lexical learning through associative processes (Bloom, 2000, p. 57).

Likewise, human translators particularly benefit from images when dealing with specialized texts (cf. Tercedor-Sánchez & Abadía-Molina, 2005; Monterde, 2004). In this sense, Kußmaul (2005) reports on some cases he studied to show the importance of images for a correct interpretation of source texts. He based his observations on Fillmore's *frame* and *scene* notions (1977).

When dealing with problems of meaning, there is no direct path from source-text word to target text-word. [...] What the translator does (or should do) is visualize a scene fitting the word (or frame, in Fillmore's terms). This scene will then [...] stimulate a target frame, that is, a translation. Kußmaul (2005)

In a related practical example, the scene or path from the technical source text to the technical target text is the image of the referenced artifact. Furthermore, in the context of online resources, a word-based image search is a useful strategy to visualize the scene when it is not familiar to the user. Once the scene or image is visualized, the understanding of the source text improves and a higher precision in the target text is achieved.

In this thesis, it is assumed that an image is the most universal representation for human beings to agree on a concept when the object itself is not present. A proof of it is the unaltered state of images during the localization process carried out on multilingual websites bearing technical content. For example, when a website is localized, i.e., translated into a different language, the text is translated and the cultural or locale specific content is rewritten and translated, but images of artifacts from the original site are kept and continue to be the object referents for terms in the new language. In this regard, Esselink (2000, p. 37) recommends building sites with directories dedicated to shared images. He mentions that only bullets, backgrounds or logos are images susceptible to be shared, but a *shared images* folder could perfectly include all the artifact images of the site. He also suggests not to include text in the images, but to store text in a different layer to make them easier to localize. If these best practices in

internationalization processes are taken into account, images will maintain their language-independent value.

Another instance of the universal nature of images in technical and scientific contexts can be observed in new generation terminology management systems (TMSs). As terminological databases are nowadays concept-oriented (see ISO TC 37/SC4 N021: 2002), TMSs support the storage of images of the concept. Thus, it is possible that a terminological record includes, for instance, three entries in different languages, with synonyms, geographical variants, acronyms, rejected terms, etc. However, it will be enabled to include just one image representing the whole concept. Such image will not change with independence of the amount of information modified, added to or withdrawn from the record.

Ortega (2002), citing Eco (1989), seems to confirm our assumption with the following statement:

The scientific image would be like a closed work as its meaning intends to be unambiguous. Its builder/emitter expects the reader to interpret the meaning of its elements as well as that of the structure that it represents in the same terms given by its creator. [our translation] (Ortega, 2002)

Fillmore's *frame* and *scene* (1977) as well as Eco's view of scientific images (1989) can be illustrated with Figure 9, which also serves to represent the two instantiations described above for the nature of images.

This illustration suggests that image and term are intrinsically indissociable. Therefore, the nature of the object in the image affects the linguistic representation and vice versa. The following section analyzes and describes the characteristics of the linguistic representation component of the BC model.

1.2.2.2. The linguistic representation component

The linguistic component of our model mainly focuses on nominal multiword terms (MWTs). For a definition of a nominal MWT, let us use two criteria from Baldwin and Kim's definition of multiword expression (MWE) and add two more criteria. Thus, we define MWTs as lexical items that: (a) can be decomposed into multiple lexemes; (b) display semantic idiomaticity (Baldwin and Kim, 2010, p. 269); (c) have a noun as their nucleus; and (d) are terminological units used in special subject fields (Cabr e, 2000).

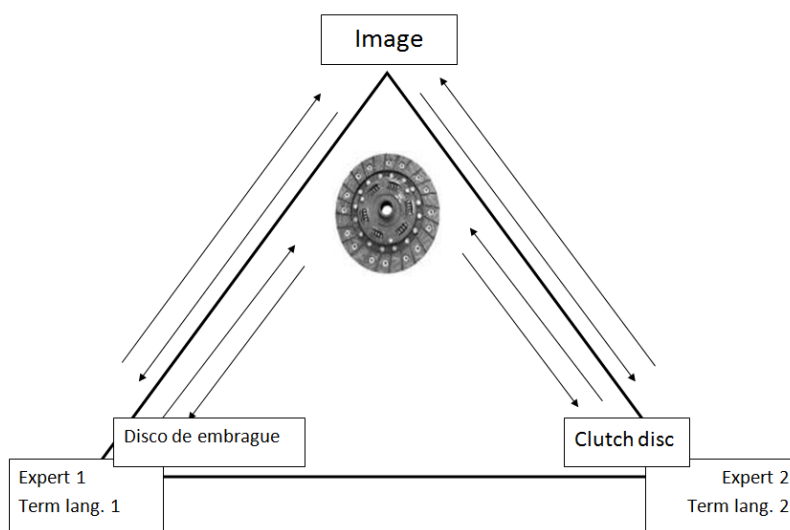


Figure 9. Image as a universal, language independent representation

Choosing nominal MWTs as the language unit for this work seems to be appropriate considering that, in specialized discourse, noun phrases are likely to be or to contain nominal MWTs. Thus, if we accept the assumptions that (a) noun phrases present a high frequency in scientific and technical discourse and (b) most terms are MWTs, it can be said that nominal MWTs show a high frequency in specialized discourse.

Likewise, nominal MWTs' adequacy lies in the fact that modification of a head noun by participles, adjectives, or other nouns makes the lexical item more specialized and less ambiguous. For instance, in the English MWT *pencil-type ignition coil*, the mere head noun *coil* has 6 different definitions in the Oxford Online Dictionary⁹. Even when some of such definitions are specialized, they continue to be of a general scope. But in the presence of the modifiers *pencil-type ignition*, the lexical item resolves possible ambiguities, increases its probability of being domain-specific and, at the same time, discards the chance to be a general language single-word lexical item.

Finally, as supportive figures in favor of the adequacy of using MWTs for this study instead of single word expressions, let us refer to dictionaries used in §1.1.1.2.3 again. Out of 46,241 entries, 36,040 (77.9%) are MWTs. On the other hand, as for the assumption that many single

⁹ <http://oxforddictionaries.com/definition/coil?q=coil> retrieved on January 29, 2012.

nouns in specialized discourse would tend to appear in general language too, the dictionary analysis shows that out of 46,241 head nouns¹⁰, 39,474 (85.5%) are included in a general language dictionary¹¹. From a practical perspective, this suggests that (a) for specialized document retrieval, single word queries would result in a big number of non-relevant documents while MWTs would retrieve more relevant ones, and (b) modification should be considered for term extraction so the list of extracted candidate terms is also relevant and reduced, which certainly would benefit further processes of term ranking or image-term alignment.

1.2.2.2.1. Premodification or postmodification?

While it is clear that Spanish nominal MWTs will mostly present postmodification rather than premodification, the difference for English is not that sharp, as stated by Biber et al. (1999, p. 578) after a corpus study:

“In all registers, noun phrases with premodifiers are somewhat more common than those with postmodifiers... Proportionally, in academic prose, almost 60% of all NP have some modifier: 25% have a premodifier; 20% have a postmodifier; and additional 12% have both.”

Therefore, postmodification was analyzed in the dictionaries introduced in §1.1.1.2.3 in order to shed some light on the representativeness of postmodifiers in nominal MWTs. The analysis showed that 33,447 (92.8%) nominal MWTs out of 36,040 did not include postmodifiers. These figures, along with the fact that Biber et al. (1999), among other authors, refer their studies to noun phrases in general rather than to nominal MWTs support our decision of considering just premodification for the English language in this work.

1.2.2.2.2. Artifact MWTs

In order to provide a more accurate characterization of the language unit for the model presented in §1.2, a further delimitation of nominal MWTs must be done. That is, out of the initial subset of nominal MWTs, just those nominal MWTs whose head nouns can be classified as **artifacts** according to WordNet’s unique beginners (Miller, 1998) will be considered. Taking

¹⁰ Not all of the entries in these dictionaries are nouns or nominal MWTs. However, the number of verbs or other type of MWTs is not representative.

¹¹ This analysis was done using a dictionary of common words as a stoplist to filter each specialized dictionary. The package is called 21dicks packages and can be reached at <http://wordlist.sourceforge.net/>.

this into account, and for the sake of terminological precision, our nominal MWE will be referred to as *artifact MWE* hereafter.

The following are some examples of artifact MWTs for English and Spanish:

English:

continuously variable cam phaser
forged-steel fully counterweighted crankshaft
single-stage turbocharger
two-stage variable intake manifold
individual throttle butterfly
pencil-type ignition coil

Spanish:

actuador eléctrico de trampillas de ventilación
termoconmutador de temperatura de refrigerante
bomba neumática de cierre centralizado
colector de admisión
llave de cincho para filtro de aceite
conjuntor-disyuntor hidroneumático de suspensión

As it was said before, artifact MWTs are taken as a point of departure in the context of specialized discourse. The fact that we work with specialized discourse means that we deal with semantic, lexical and morphosyntactic features according to very precise pragmatic and communicative situations in domain-specific documents, e.g., automotive engineering. Such situations will range, for example, from product catalogues published by manufacturers for distributors (i.e., expert to expert) to technical manuals or product catalogues developed by distributors for clients or end users (i.e., expert to expert or expert to non-expert). This is the reason why we present the following considerations from the perspective of terminology.

1.2.2.2.3. Terminological basis

The nature of the search space potentially containing term translations, that is, automotive engineering digital documentation, allows for a description of the problems posed by artifact MWTs. Such are the challenges intended to be solved by proving our hypothesis (see §1.3) according to the below postulates proposed by Cabré (2000):

1. As for the terminological object:

In this sense, we consider that the object of study of terminology as a discipline are the terminological units used in special subject fields, and that these units have to be analysed functionally, formally and

semantically by a description of their dual systematic nature: i.e. their general systematic nature in relation to the system of the language to which they belong; and their specific systematic nature in relation to the terminology of the domain in which they are used. (Cabré, 2000)

2. As for the terminological variation principle:

Any process of communication involves variation of lexical forms, which manifest themselves as alternative denominations for the same concept (synonymy) or in the semantic openness of one form (polysemy). This principle applies to all terminological units, although in different degrees, according to the type of communicative situation. The greatest degree of variation occurs in discourse destined to popularise science and technology; the smallest degree of variation is characteristic of terminology standardised by groups of experts; a middle position is characteristic of the terminology used among specialists in everyday communication. (Cabré, 2000)

The phenomenon of variation is one of the facts driving the present research. From this perspective, our main search space for term translations will be documents of high specialization and medium variation as well as documents of medium specialization and high variation.

In a related study of knowledge representations in the field of aeronautics, Monterde (2002, p. 221-223 and 2004) concluded that photographs are object representations just present in text of low specialization. However, observations done so far for the present research in web sites as well as in catalogues and other hard copy material suggest the opposite. There are three elements which could account for this discrepancy with Monterde's view: a) the target audience, b) the degree of consolidation of a subject-field and c) the specificity of the information. For example, a catalogue targeting a distributor of industrial automatic devices will necessarily include photographs of devices and spare parts which certainly will be out of reach of lay people for a correct recognition and interpretation.

Catalogs and manuals certainly represent a particular genre, but they still are specialized texts with terminological variation which constitute an interesting search space for the location of term translations with object representations such as a photograph. In this line, Gamero (2001, p. 73) observed that manuals, spare part lists and technical descriptions are expositive texts which use to have illustrations whose language referents can be found in surrounding text. She also says that this kind of texts is addressed to specialized receivers.

In the previous sections, the characterization of the linguistic and visual representation components of the BC model has been carried out. This makes it possible to sketch the general

methodology to implement a recursive and automated process for image-term and terminology alignment. The methodology is presented below.

1.2.3. Implementing the BC hypothesis

For the implementation of the BC, the image and term retrieval tasks will be automated. Image retrieval is carried out by means of content-based image retrieval (CBIR) techniques, and terms are retrieved with support of term recognition and noun classification techniques. They are briefly described below.

CBIR is a visual information retrieval method which uses no linguistic information but images as examples to retrieve similar images from an image set. Main color, shape or texture features are extracted from the example image and then compared to the same features of each image in the image set. Images with similarity or distance values below or over a given threshold are considered of the same class, that is, similar.

As it is described in Chapter 2, despite the existence of a number of tools, our main problem for CBIR remained unsolved. Some tools were not available, some others were closed demos, some other were too cryptic to be used or installed, and the others did not perform well when tested with our prototypical image. We needed therefore a tool that maximized his performance with the type of images that usually co-occur with the linguistic component of the BC model, that is, with a multi-word term. This is how and why DORIS (a Domain-ORiented Image Searcher) was born. DORIS (Jaramillo and Branch, 2009) is the product of a collaborative work between researchers of the National University of Colombia and the author of this thesis and is tailored to the relevant type of images for the present research. DORIS is integrated in the BC model and implemented in a prototype for term translation and image annotation described in Chapter 5. An evaluation of its performance is also presented in the same chapter.

With regard to term recognition and noun classification techniques, some low level tasks such as part-of-speech tagging and chunking are used. Some high level processing is also done to assign lexical semantic classes to nouns and classify them into concrete and abstract. For these tasks, available tools are used.

Figure 10 shows the automation proposal for term alignment and image indexing. It explains how two documents that are theoretically interrelated by the BC hypothesis can be connected in practice. A spider is launched to the Web. Websites fulfilling predefined criteria (e.g., in a Web directory) are saved and their images analyzed and indexed by DORIS. If an image in the database presents feature values within a threshold determined by the example image features, nouns in its website are classified and extracted from the surrounding text to make up a list of candidate target terms which could designate the object in the website's image. Last, target image and target term are aligned, and then source term and target term are aligned too.

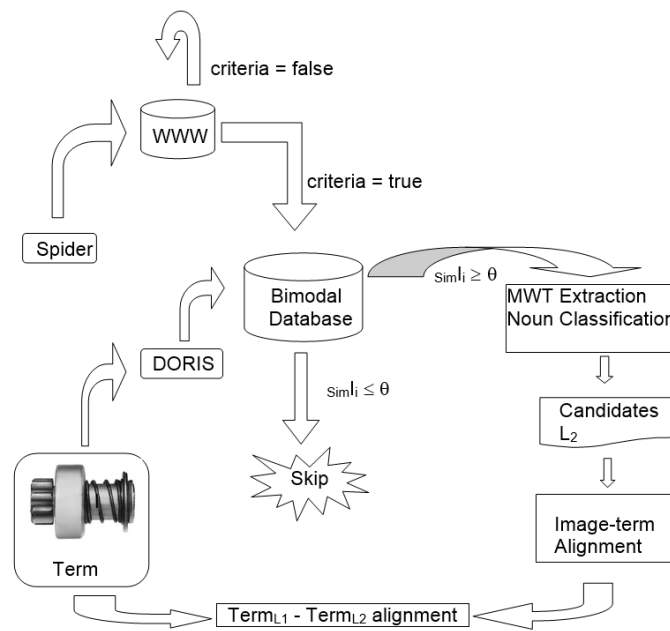


Figure 10. BC-based methodology for term alignment and image indexing

1.3. Assumptions and hypothesis

1.3.1. Assumptions

The following assumptions underlay the development of the BC model:

- a) The co-occurrence of images and their index terms in technical and scientific documents is a language- and domain-independent phenomenon, although it can be more common in some languages and in some domains.

- b) Terms are susceptible of variation as a result of translation or direct denomination processes. Likewise, such variation can be found in monolingual contexts as well as in multilingual ones.
- c) Artifacts can be represented by images and by artifact (concrete) nouns; both representations are naturally used in specialized technical documentation and, when both modes are related, they constitute a bimodal co-occurrence.

1.3.2. Hypothesis

In this thesis, it is assumed that the bimodal co-occurrence of images and terms is natural to any discourse, in a greater or lesser extent, according to aspects which are inherent to each language's socio-cultural and/or technological resources. In the very moment that such co-occurrence simultaneously exists in two documents of different languages (or even the same language) for an identical artifact referent, it can be stated that both images as well as both linguistic denominations designate the same artifact and, therefore, the terms are equivalent. Accordingly, equivalent multilingual terms can be located and retrieved by initially matching object representations (i.e., images) of artifacts. The nature of this interface, allows for the retrieval of equivalents with independence of their morphological, syntactic or lexical configuration.

With this hypothesis as the basis of this thesis' methodology and products, we expect over-generation of terminological variants to be reduced. The reduction will be achieved in that the model will enable its users to take advantage of existent variants for technical translation and writing. On the other hand, the fact of using a term found elsewhere assures a minimum of usage convention among a community of users. Thus, texts or translations generated with support of this model would probably share common terms with other documents which will also have positive consequences in information retrieval.

It is also expected that terms (in context) generated by the model help users deciding on the appropriateness of the term for a specific task. In this scenario, visual dictionaries as products of the model will play an important role too.

Finally, dictionaries or term bases compiled with the proposed model would document terms with a longer average length than traditional resources. This will be achieved thanks to image-term alignment processes which will match images with complete multi-word terms.

1.4. Objectives and delimitation of scope

1.4.1. General Objective

The general objective of this thesis is to contribute to the compilation of wide-coverage dynamic multimodal terminological resources as raw material for translation- and terminology-based tasks.

1.4.2. Specific objectives

- ◆ To propose a theoretical model and a practical implementation of a concept-based strategy to dynamically compile wide-coverage bimodal terminological dictionaries.
- ◆ To establish a bimodal co-occurrence model to align term and image.
- ◆ To analyze and characterize the search space which hypothetically yields terminology to be used for dictionary compilation.
- ◆ To identify, propose and integrate techniques and tools which procedurally concretize the bimodal co-occurrence model.
- ◆ To evaluate the bimodal co-occurrence model for the specific task of building up bimodal terminological dictionaries.
- ◆ To design a functional prototype for the practical implementation of the BC model.

1.4.3. Delimitation of scope

Besides the scope stated in §1.4.1 and §1.4.2, some other boundaries for the present work are:

- a. This thesis mainly attempts to contribute to the efforts of term-image association and artifact noun classification in the context of online or local catalogues or databases. The success of these attempts will contribute to the general and specific objectives at a relative extent.

- b. Special attention will be paid to any useful by-product derived from this work (e.g., a method to produce comparable corpora). However, for the sake of relevance and priorities, the generation of any by-product will remain peripheral.
- c. In spite of the importance of images in this proposal, this thesis does not intend to focus on image retrieval or image processing. The efforts done by other researchers in that direction have been capitalized in order to make the most of the bimodal co-occurrence. That is why this thesis will mainly focus on text processing.
- d. Whenever images are mentioned for the goals of this research, it should be noted that we are referring to photographic images of artifacts. This decision was made because a) catalogues and databases mainly contain photographs, and b) the software used to match images yields better results when dealing with photographs than with illustrations or diagrams, for example. We are aware, however, of the importance of illustrations in fields such as patent licensing (cf. Donnell, 2005) which can constitute an open line for future research.
- e. English and Spanish were defined as the working languages for this research. Nevertheless, a certain degree of language and domain independence is suggested and justified.

1.5. Fields of application

From the methodology outline, different fields of application can be inferred:

- a. *Terminology*: For terminological glossary compilation which also includes visual information.
- b. *Translation*: For location and retrieval of equivalents and its variants assuring a minimal usage by experts of the subject-field.
- c. *Machine translation*: For location of parallel and comparable corpora which serve as input for alignment tools and statistical machine translation.

- d. *Language for specific purposes and corpus linguistics*: For compilation of parallel and comparable corpora which serve as frame for comparative analysis of specific language units.

1.6. The Structure of the Thesis

This thesis is structured as follows. Chapter 2 presents the state-of-the-art of work and techniques directly or indirectly related to the present work. The main emphasis will be on image-term alignment, content-based image retrieval, cross-language image retrieval, domain-specific cross-language information retrieval, term recognition, and noun classification.

Chapter 3 characterizes the search space containing instances of the BC. Some figures are given to show the worthiness of the web for the specific purpose of this research although drawbacks are mentioned too. The criteria for the search space definition are set in each language. Then a web segment is selected as our search space, its main characteristics are analyzed and data are provided. The methodology for web segment crawling is outlined.

In Chapter 4, we propose strategies towards image-term alignment, using first multi-word term (MWT) recognition techniques and then artifact noun classification. MWTs constitute term candidates for an object which is represented in an image within a given context. Illustrative examples of main scenarios are provided, that is, whether there is a lot of text surrounding the image or whether there is scarce text or just the term. Appropriate methods for MWT recognition are implemented according to the literature review in Chapter 2. Likewise, three approaches for artifact noun classification are presented. The first approach uses non-linguistic variables, the second is based on linguistic patterns and the third uses lexical semantics information taken from WordNet (Fellbaum, 1998) and EuroWordNet (Vossen, 1998).

Chapter 5 describes a prototype that integrates the different phases outlined so far and puts the BC model into practice. This chapter also provides an overall assessment of the image matching algorithm used for the present research. In this same line, as CBIR techniques certainly are not 100% effective, a prototype image must be selected for testing. It means that in order to prove this research validity, it should suffice the adequacy of the approach with instances of a prototypical image. If such success is achieved, the fact of automatically matching more complex images is a matter of refining CBIR techniques and it should not affect this research hypotheses. Therefore, a prototypical image is defined according to identified features

which, in turn, could have interesting consequences in the linguistic representation of the depicted artifact. A CBIR software is selected and a representative number of queries is made in order to have enough data for evaluation and prediction.

Last, Chapter 6 concludes with the main remarks and inferences on the major phases of this work. We show here that our central hypothesis was proved and that the general and specific objectives were accomplished. This chapter also discusses limitations and new lines for future research based on the present work's findings and open topics.

2. STATE-OF-THE-ART

2.1. Introduction

In this chapter, the present research is contextualized by tracing relevant previous works related to the core of this thesis. We address approaches and issues related to the whole bimodal co-occurrence model as it is defined and described in the Chapter 1 for image-based term retrieval (§1.2). This includes a discussion on image-term alignment, content-based image retrieval (CBIR), cross-language image retrieval (CLImR), and domain-specific cross-language information retrieval (CLIR).

Later, we focus on the literature related to the tasks, processes or methods that are necessary for the practical implementation of the BC model. We survey some of the representative works on each relevant task and examine them in the light of our problem, data, and needs. Such examination may lead to the use and/or adaptations of existent methods, although they may not constitute new proposals by themselves. In this part, we scrutinize and contextualize works on multi-word term (MWT) recognition and noun classification, as they present a direct relation with the tasks addressed in Chapter 4.

2.2. Image- and text-based information retrieval systems

Multimodality, or the interaction of different modes of communication that takes place in a specific context, became particularly attractive some years ago with the advent of the internet. Digital words, images and sounds supporting and complementing each other in their communicative purpose made researchers think about potential ways to exploit and automate such interactions. In this context, the interest of the present research was born from the implications of the link between linguistic and visual representations in specialized documents in the context of the World Wide Web. That is, a bimodal co-occurrence whose components are described in Chapter 1, and which gave rise to the proposal of a bimodal co-occurrence (BC) model presented and defined in the same chapter. Let us remind here that the overall goal of the BC model consists of using images to retrieve terminology from the web. At the time of the first explorations of this research, there was no evidence in the literature of an approach

with the characteristics and purpose of the BC model. And there are not any either nowadays. There are, however, related approaches which served as inspiration for our proposal and as a point of departure for the adaptation or creation of the necessary methods and techniques.

The downside of the existent approaches with regard to the interest of this research has to do mainly with their domain of application and scope. The fact that most of the reported work addresses images and/or documents of a general interest determines the difference in the methods and techniques used and, therefore, in the performance when tested in specialized domains. For example, Tollmar et al. (2004) report the design of an image-based system called IDEixis that aims at retrieving documents from the Web. It relies upon a closed server-hosted image database for the retrieval process. The system works through a mobile phone with a camera. The user takes a photograph of a place and sends it to IDEixis which in turn uses content-based image retrieval techniques (CBIR) to compare the photograph with its keyword-indexed image database. If there is a positive match, that is, if the database contains any image similar to the user's photograph, the system queries a search engine (Google, in this case) with the keyword(s) associated to the database image extracted with a simple but practical frequency-based keyword extraction so the user can confirm the relevance of the results or use them for new searches. The search engine returns Web pages textually- and visually-related to the keyword(s) and, therefore, to the user's photograph. IDEixis compares the images in the Web page with its image database and provides the end user with the most relevant Web pages.

For its practical application in the tourism field and considering the rapid progress of mobile devices, IDEixis is an interesting proposal. Unfortunately, the evaluation of the system is only given in percentages, which provides a limited view of its real performance, and it seems not to be available in the market or for public use and evaluation either. It can be observed, though, that it is an application designed for tourists. Therefore, its image database contains pictures of famous places, and its image and text processing strategies point to documents of general interest. This particularity, as discussed later in this section, makes it inappropriate for its application in the search space described in Chapter 3 given the formal and semantic characteristics of the constituents of our BC hypothesis, that is, our prototypical image (Chapter 5, §5.5.2.5) and relevant linguistic representation(Chapter 1, §1.2.2.2).

The same drawbacks related to such general scope apply to other more recent implementations. Google, for instance, added content-based image retrieval (CBIR)

capabilities to its traditional text-based image search engine¹². But, the fact of indexing millions of images of different types makes the image set highly heterogeneous. Therefore, accuracy tends to decrease or to depend too much on a number of factors. Indeed, Google maintains: “...you’ll likely get more relevant results for famous landmarks or paintings than you will for more personal images like your toddler’s latest finger painting.” With these constraints, there are very few chances of obtaining good results for images that respond to particular needs. In fact, we checked its performance with a small number of domain-specific image queries (e.g., a *diode rectifier*) that yielded no positive or even fuzzy matches.

Given the small number and the limitations of available systems for effective bimodal information retrieval, it is necessary to break down the state-of-the-art review into the central components and functionalities of the BC model. Let us now start by presenting and discussing the image-term alignment-related contributions.

2.2.1. Image-term alignment

The image-term alignment problem gains importance in settings where an image is surrounded by any amount of text. The goal of the task in this specific setting is the alignment of the image with the linguistic expression in its surroundings that best denotes the object(s) in the image.

The image-term alignment problem is different from that of image annotation. The latter has received great attention and discussion, while the former seems to be less explored. Image annotation consists of classifying an image under a predefined category (*landscape, animal, building, etc.*), while image-term alignment seeks to find the image description in collateral information. A good deal of image annotation works have been presented in the cross-language image retrieval track of the CLEF initiative (ImageCLEF¹³). ImageCLEF also features a task in cross-language image retrieval (CLImR). The relevant approaches for CLImR are being presented later in this chapter. (§2.2.3). Suffice here to remind that CLImR, as traditionally tackled, aims at retrieving images from collections whose text is in a language different from that of the query. The goal, therefore, is to retrieve relevant images but not to align images with its best textual descriptor in the surrounding text. A detailed description of the tasks, approaches, data sets, and results of recent image retrieval and image annotation tasks can be

¹² <http://www.google.com/insidesearch/features/images/searchbyimage.html>

¹³ <http://www.imageclef.org/>

found at Tsirikika et al. (2011) and Villegas and Paredes (2012). Even though image annotation could interestingly support the image-term alignment problem, we do not dig here into the former but focus more on the latter.

One of the earliest efforts to associate text to images was reported by Srihari and Burhans (1994). The authors analyze newspaper image captions and attempt to predict which objects are present in the image and to generate constraints useful in identifying faces in the picture. They use PICTION, a tool they authored to identify faces using information in caption's textual descriptions about visual features in the image. PICTION does not detect faces but rather locate them in the image according to key expressions in the descriptions. Captions are semantically and syntactically parsed with support of dictionaries, lexical knowledge bases and manually characterized visual information. The authors report 62% of correctly identified faces in an initial evaluation of PICTION.

Paek et al. (1999) also experimented with news articles in order to classify indoor and outdoor scenes using collateral text. For text-based classification of images, they used the *tf.idf* measure for the first sentence of captions. For images, an analogous *of.ijf* (object frequency and inverse image frequency) through object recognition and segmentation was innovated. In addition to this, the authors also implemented a machine learning technique. For this purpose, images were manually annotated and used to train the statistical model. Classification accuracy with a combined approach of text and image features is 86.2%.

In a different domain, Ahmad et al. (2002) propose a modular approach through the use of several neural networks for classification of visual features, textual features, and the connections between both sets of features. They test the networks on crime scene images annotated by experts with specialized short descriptions –10 words in average. To build image vectors, low level features are computed. For text vectors stopwords are removed and relative frequencies of remaining content words are drawn from specialized texts on the one hand, and from a general purpose corpus on the other. Two networks take a vector each and contribute to a final decision in a sort of voting system. A third network links the winning neurons of the image and text input and yields a final decision on what are the appropriate textual descriptors for the image. The authors report 74% of accuracy for text classification and 48% for image classification.

News articles are again used by Feng and Lapata (2008). The authors combine text and image features to train a relevance model adapted from Lavrenko et al. (2003) that captures the joint probability of images and annotated words directly, without requiring an intermediate clustering stage. Tags for images are content words extracted from captions and the main article. Images are not properly segmented but analyzed using grids to extract salient features from rectangular regions. The authors found that combining image, caption, and the whole article improves the overall performance which is measured using precision and recall. They report an F1 score of 19.82.

One of the most recent works on annotation of images with information from collateral text has been reported by Leong et al. (2010). The authors do not use image features, but they rather completely rely on textual information. Their method was tested on web documents with unrestricted text. The test set consisted of 300 image-text pairs which was built using the British National Corpus most frequent words to retrieve specific size images from the web having at least 10 surrounding words. Each image was manually tagged with content words (mainly nouns). Three unsupervised methods were tested and then one supervised method combined them all in a voting system. Their methods are: 1) Flickr is used to extract labels related to a set of source words, assign weights according to the degree or relatedness of the source words with Flickr tags, and thus measure *picturability* of these words so the ones with higher weight are assigned as tags for the source image; 2) a graph-based keyword extraction method enhanced with a semantic similarity measure was used to extract keywords from Wikipedia; the output is a sorted list of words in decreasing order of their ranks which are used as candidate labels to annotate the image; 3) a third method is called topical modeling. They use the Pachinko Allocation Model (PAM) (Li and McCallum, 2006) to characterize the topics in a text. This is a sort of document classification. The topics yielded by this method are potential labels for an image. The authors compute precision and recall for each individual method, but also for a supervised method (using support vector –SVM– machines) that combines the results of the three unsupervised tagging classifiers. The SVM reaches overall better scores than individual methods. Their combined SVM method not only outdoes the baselines, but also performs statistically as well as the textual method reported by Feng and Lapata (2008), which seems to confirm the power of text-based approaches.

Last but not least, it is worth mentioning some of the interesting work that has been undertaken in the domain of patent search and retrieval. Vrochidis et al. (2012), for instance, propose a method for concept-based patent retrieval. The authors extract textual features as well as low-level visual features which are used to build vectors for further classification with SVM. The dataset was manually extracted from 300 patents out of which 1,042 images were annotated by humans. For image annotation and for text feature extraction, image descriptions, i.e., captions, were used. As images are black and white and depict technical information in a diagrammatic form, Adaptive Hierarchical Density Histograms (ADHD) are used to map image features. Captions, on the other hand, were modeled using the bag of words approach. Frequency was computed for all the captions and words in individual captions were assigned weights when they appeared in the frequency list, which was made up by 100 words. Therefore, 100-dimension vectors were build. Three classifiers were trained with the generated features, namely, one with visual features, another one with textual features, and a third one with a combination of both. A cross-validation technique was applied, and precision, recall, and F score were computed for evaluation. The overall F scores by the different classifiers were: visual: 63.88, textual: 75.35%, and hybrid: 78.95%

The literature review suggests that the use of collateral information for image annotation has been an area of interest for several years. It can be noticed, however, that most of the work has been addressed to news articles or general interest documents. The only attempts to tackle the problem in a specialized domain were reported first by Ahmad et al. (2002) who even propose a measure for termhood that they call *weirdness*, and later by Vrochidis et al. (2012). With the exception of these studies, a shortage of research is evident when it comes to the problem of texts and images in specialized domains.

It is precisely the use of specialized text and images that adds particular challenges to the image-term alignment task. As for image features, focus on general purpose images makes any sort of segmentation or image modeling necessary, as reported by most of the works –with the exception of Leong et al (2012). Working with artifact images as the one characterized in Chapter 5 (§5.5.2.5) certainly poses interesting problems, but discards the segmentation problem.

With regard to processing of collateral text, the domain also has practical and conceptual implications. One of the most remarkable differences imposed by a specialized domain is the

length and the terminological nature of tags. As most of the cited works focus on general purpose documents, tags assigned to images tend to be general languages single lexical units. In the present thesis, however, we deal with multi-word terms (MWTs) as tags, which implies the additional problem of MWT identification, extraction, and alignment. Certainly, general language also features multi-word expressions (MWEs). Therefore, the fact that single-word expressions are most of the times used as tags during the alignment process suggests the idea that truncated MWEs are many times used as tags. It is the case of one of the examples cited by Feng and Lapata (2008) where the original caption is “*Thousands of Israeli troops are in Lebanon as the ceasefire begins.*” and the proposed tags are “*Lebanon, Israeli, Lebanese, aeroplane, troop, Hezbollah, Israel, force, ceasefire, grey.*” In this example, the noun phrase *Israeli troops* was split and its constituents computed as individual tags.

Besides these specific issues, from the previous discussion it can be inferred that content-based image retrieval (CBIR) is a core component of an image-based information retrieval system. And so it is for our BC model. Even though we do not rely on CBIR for the image term alignment task, we do use it as a pivot for location of target documents and terms (see Chapter 1, §1.2). While image processing is out of the scope of this work, the next subsection briefly presents a definition of what CBIR is and provides some insights into the existent works and their relevance for the present research.

2.2.2. Content-based image retrieval

Content-based image retrieval (CBIR) technology is part of the visual information retrieval frame. It aims at retrieving still images according to some specified features which are defined from predetermined parameters or from a query image whose attributes are automatically extracted. As for the use of parameters, the objective is to retrieve images by means of the characterization of features potentially present in an image i_i from an image repository H , so that the system returns only images containing the specified features. In this case, the only reference for relevance checking in the results is the existence of the initially required features in the returned images. A typical query for this approach could be, for example: *I'm looking for images with 60% of green, 20% of blue, 10% of white and 10% of brown.* This query would probably return images of landscapes. Therefore, such approach will be particularly useful when trying to retrieve general-feature images, e.g. any landscape, any flower or any airplane, but not when searching specifically for, say, an image of the Mount Everest, a daisy or an F-16 aircraft.

The latter problem of retrieving images with more specific features is addressed by querying the image repository H with an example image i in order to retrieve similar images from H . In this scenario, the required features to be searched in i are automatically extracted from i . According to such features, a threshold is defined to determine the degree of similarity between i and i . Some of these image features – which are called low level features – are texture, shape and color as well as a combination of them. The most relevant image out of the retrieved results will be the one presenting the greatest similarity with i . Some variations such as image translation, rotation, scaling, deformation or a combination of them have also been taken into account to improve precision when designing CBIR applications (see Figure 11).

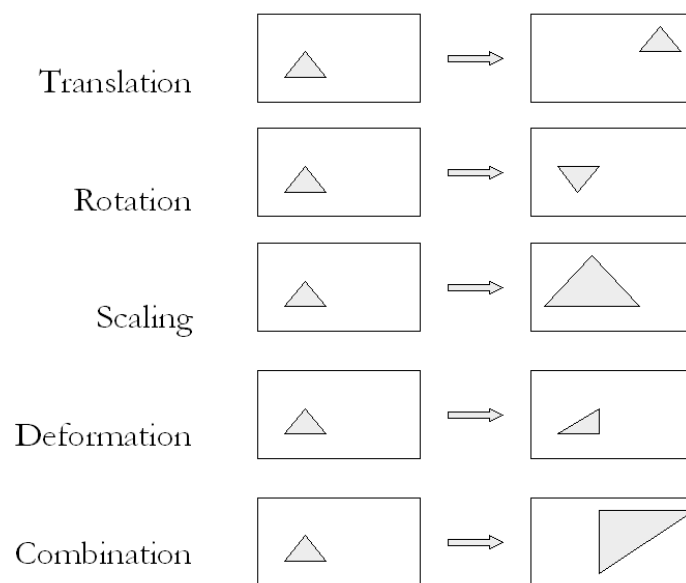


Figure 11. An Illustration of image translation, rotation, scaling and deformation. The last frame shows a combination of all of them.

While the approach is the same for most of the proposals reported in the literature, what does change are the descriptors, the algorithms and the mathematical methods used by the different systems. As for descriptors, the Scale-Invariant Feature Transform (SIFT) (Lowe 1999) is one of the hottest ones in computer vision applications. Other examples of well-known sets of descriptors are Gradient Location and Orientation Histogram (GLOH) (Mikolajczyk and Schmid 2005), Speeded Up Robust Features (SURF) (Bay, et al. 2008), the machine-optimized gradient-based descriptors (Winder and Brown 2007), (Winder, et al. 2009) and the well-established MPEG-7 descriptors (Manjunath, et al. 2002).

Algorithms are designed and descriptors are selected according to the most prominent features in the image data sets. As previously mentioned, CBIR efforts have focused so far on general image collections, that is, images of people, places, or landscapes. In one word, very noisy images. Noise, in this context, means that the object of interest in the image cannot be clearly differentiated from the background or from other objects, but it is surrounded by other irrelevant objects (e.g., trees, other people, buildings, cars, etc.). This makes the task more difficult and compels further preprocessing of the image to attain prior object segmentation. Of course, the more complex the image, the more preprocessing is needed, and the more inaccurate the final outcome.¹⁴

This seems, then, the standard approach followed by most of the CBIR systems, e.g., CIRES¹⁵ (Iqbal and Aggarwal, 2003), QBIC¹⁶ (Flickner *et al.*, 1995), PHOTOBOK¹⁷ (Pentland *et al.*, 1996), and VisualSEEk¹⁸ (Smith and Chang, 1996). There are some more sophisticated commercial applications such as IMATCH¹⁹ that provides an ampler range of setting options. The system comes preconfigured with certain values and then the user can modify those values according to the features of the image to be retrieved from a database. Such customized configuration, however, adjusts to a specific query and, unless all the images share very similar features, the parameters have to be set every single time.

With the intention of improving performance and overcoming the inaccuracy of the initial results, some have proposed queries with groups of example images instead of a single example image. When this group of images is the result of the first query performed by the user, the procedure is called relevance feedback. It enables the user to interact with the system by selecting the most relevant images from the retrieval result (cf. Nakazato *et al.*, 2003 and Iqbal and Aggarwal, 2003). Sometimes, it is the system that automatically takes the k -first results to

¹⁴ For a detailed description of the CBIR standard technology, the reader can see Urcid Pliego (2003) or Geradts (2003). Rui *et al.* (1999) also present concrete information on the main features for CBIR as well as on some related systems and research. And an updated review, compilation of CBIR techniques, real world applications, evaluation techniques and interesting additional references can be found in Datta *et al.* (2008).

¹⁵ <http://amazon.ece.utexas.edu/~qasim/research.htm>

¹⁶ <http://www.research.ibm.com/topics/popups/deep/manage/html/qbic.html>

¹⁷ <http://vismod.media.mit.edu/vismod/demos/photobook/>

¹⁸ <http://www.ee.columbia.edu/ln/dvmm/researchProjects/MultimediaIndexing/VisualSEEk/VisualSEEk.htm>

¹⁹ <http://www.photools.com/>

extract features and build a new query; this has been called pseudo-relevance feedback or blind relevance feedback.

With some exceptions, available tools generally retrieve images from a closed local database, which does not help much when the image set in the database is irrelevant for the user. For instance, it is useless to query a wildlife database with an image of an artifact. Integrating the option to create and query a new relevant database becomes then an important functionality. This limitation is what makes some of the applications that have been implemented on the web merely interesting but limited demos. A few proposals seem to go beyond and directly interact with the Web although its restrictions remain the same, that is, they address general images and general language, and some are not available for use or testing. Webscope (Yi *et al.*, 2000), for instance, grabs images from the Web, stores them in a database and extracts visual and semantic features (through html meta-tags) which are then combined to improve retrieval precision. Such semantic features are basically categories assigned to directories or files names in the file structure of the server hosting the image.

Webseek²⁰ (Chang *et al.*, 1997) is an online system that had indexed 650,000 images and 10,000 videos up to the publication date. The indexation process for most of this visual material was carried out by Web spiders during three months. Based on a visual and textual index (visual by color binary sets and color histograms, and textual by text and html tags surrounding the image), Webseek retrieves images faster. However, in spite of the fact that CBIR technology has been added to the system, users had queried it with categories 53.5% of the times and just 3.7% of the times they have use the system's CBIR capabilities²¹. This suggests little familiarity with the approach by the user, preference for the category method, or dissatisfaction with the query by image results.

Cortina²² (Gelasca *et al.*, 2007) uses a Web crawler and starts from the Open Directory Project (ODP)²³ categories to grab images from the Web which are then analyzed and indexed. By the time of the referred publication, Cortina reported about 10 million indexed images. Besides the

²⁰ <http://www.ee.columbia.edu/ln/dvmm/researchProjects/MultimediaIndexing/WebSEEK/WebSEEK.htm>

²¹ The publication is not clear with regard to the remaining percentages.

²² <http://vision.ece.ucsb.edu/multimedia/cortina.shtml>

²³ <http://www.dmoz.org/>

ODP-based automatic annotation, images can be manually annotated by users through its Web interface. It seems that users were previously allowed to upload their own images for further retrieval, it seems such capability is not enabled anymore and only a registration functionality for aerial images is currently available.

As for more recently available Web-based tools, let us mention TinEye²⁴ and I-Search²⁵. TinEye “looks for the specific image you uploaded, not the content of the image. TinEye does not identify people or objects in an image... It finds out where an image came from, how it is being used, if modified versions of the image exist, or if there is a higher resolution version.” It means that in order for the search to be successful, the query image has to be taken from a website and the website has to be previously indexed. I-Search, on its part, is a multimodal search engine that supports audio, video, rhythm, image, 3D object, sketch, emotion, social signals, geolocation, and text (Axenopoulos et al., 2012). It has a furniture and ethnomusicology database, and the user is allowed to upload media to query the database, but there is no capability to build or index a personalized database.

The exploration of available tools also let us discover some downloadable freeware for CBIR. Some of them widely used, so to speak, as is the case of GIFT²⁶, and some others based on reasonable solid research and experimentation such as Blobworld (Carson *et al.*, 2002) and ImageGrouper²⁷ (Nakazato et al., 2003). They are useful tools in the sense that they contribute to handle the overall problem of CBIR, but also because they enable users to create and index their own databases as well as to customize certain parameters.

2.2.3. Cross-language image retrieval

As stated in Chapter 1 (§1.2.1), the phenomena represented in the BC model can also take place in bilingual settings. The problem of finding information in document collections of a different language falls on the grounds of cross-language information retrieval (CLIR). But if the goal is to retrieve images in document collections with text in other languages, the problem is addressed by cross-language image retrieval (hereafter CLImR). While domain-specific CLIR

²⁴ <http://www.tineye.com/>

²⁵ <http://vcl.itl.gr/is/isearch/client/musebag/>

²⁶ <http://www.gnu.org/software/gift/gift.html>

²⁷ <http://otm.illinois.edu/technologies/imagegrouper>

is briefly discussed later in this chapter, this subsection is devoted to review the general approaches used to find relevant images in multilingual collections in the frame of CLImR.

CLImR focuses on the linguistic content associated to images so that a collection in the target language can be queried through keywords in the source language. The most common approaches in the literature are a mixing of CBIR and CLIR. This happens, in part, because the methodology for CLImR shares many features with that followed for CLIR considering that there is always text associated to images. This in a typical CLIR system can be summarized as follows: a) a user poses a query; b) often, an automatic query expansion is done by adding more words to the query in order to improve the retrieval performance; c) query normalization takes place; d) the query is translated into the target language(s) or the documents in the target language(s) are translated into the query language; e) once the query and the document collection are in the same language, a monolingual information retrieval takes place; f) retrieved documents are ranked according to their relevance with respect to the query and presented to the user.

But these mixed approaches also occur because one technique (CLIR or CBIR) sometimes does not yield the best results alone. It is the case of the study reported by Besançon et al. (2005) who first try a simple approach to merge the scores of image and text search. Then, in a first experiment, they give more weight to text retrieval and in a second experiment image retrieval is given more weight. When a specific modality is given more weight, the other one is mainly used to reinforce the search results of the first modality. They performed experiments on two different image collections which yielded divergent results because of the image and text corpus characteristics. This reinforcing merging strategy slightly improved retrieval precision (cf. Besançon and Millet, 2006).

The mixed approach is also beneficial for Alvarez et al. (2005) and Chang and Chen (2006). Alvarez et al. (2005) first perform a keyword-based image search. Then, image processing is carried out on retrieved images to identify the class of visual features (e.g., texture, shape, color) that is mostly associated to a specific keyword. For example, “the keyword *animal* could belong to the *shape* class since the measure using shape information will be the most discriminant to identify images with animals (although *zebra* and *tiger* will probably belong to the *edge* and *texture* classes)”. Then they use such classes of visual features to retrieve similar images by content. Chang and Chen (2006) automatically populated a WordNet-based ontology with images in

order to obtain a sort of intermedia. An example image was used to query this ontology and the concepts associated to the retrieved images were used to carry out a text-based search in the ontology again with the purpose of retrieving concept-associated images. An additional experiment was undertaken directly on the IAPR TC-12 Benchmark used as intermedia. The results of image- and text-based merging for this collection outperformed the experiments with the ontology.

There are, however, other cases where a mixed approach lowers accuracy. Daumke et al. (2006), for instance, used their CLIR subword approach for medical image retrieval. They first normalize the image annotations, convert them into subwords –defined as self-contained, semantically minimal units– and take them to an *ad hoc* interlingua. The conversion of annotations into subwords is carried out against a manually built lexicon and thesaurus of biomedical terms. Their experiments yielded satisfactory results for textual retrieval but poor results for mixed textual and visual retrieval.

A comparison of the best mean average precision (MAP) scores of some tasks at the ImageCLEF²⁸ track in Tables 4 and 5 shows how the interaction of visual and mixed features may affect precision.

Task	Modality	Best MAP
Medical image retrieval task (Müller et al., 2007)	Textual	0.3962
Wikipedia image retrieval task (Tsikrika et al., 2011)	Mixed	0.3880
Photographic retrieval task (Clough et al., 2006)	Mixed	0.385
Photographic retrieval task (Grubinger et al., 2007)	Mixed	0.3175
Medical image retrieval task (Müller et al., 2008)	Mixed/Text	0.29
Wikipedia image retrieval task (Popescu et al., 2010)	Mixed	0.2765
Wikipedia image retrieval task (Tsikrika et al., 2009)	Textual	0.2397

Table 4. Best MAP scores for some ImageCLEF tasks - Different modalities.

Task	Modality	Best MAP
Medical image retrieval task (Müller et al., 2007)	Visual	0.2328
Photographic retrieval task (Grubinger et al., 2007)	Visual	0.1890
Photographic retrieval task (Clough et al., 2006)	Visual	0.1010
Wikipedia image retrieval task (Popescu et al., 2010)	Visual	0.0553
Medical image retrieval task (Müller et al., 2008)	Visual	0.04
Medical image retrieval task (Müller et al., 2010)	Visual	0.0358
Wikipedia image retrieval task (Tsikrika et al., 2009)	Visual	0.0079
Wikipedia image retrieval task (Tsikrika et al., 2011)	Visual	0.0044

Table 5. Best MAP scores for some ImageCLEF tasks - Visual modality

²⁸ <http://www.imageclef.org/>

It can be seen in this comparison that textual retrieval outperforms mixed and visual retrieval. The poor performance of the visual retrieval runs results from the complex and heterogeneous image datasets. For this specific evaluation campaign, two collections with similar characteristics have been used as benchmarks for CLImR evaluation: first, the black-and-white St. Andrews collection of historic photographs²⁹ and later the IAPR TC-12 Benchmark. The St. Andrews collection consisted of 28,133 images, all of which have associated textual captions written in British English. The captions consist of 8 fields including title, photographer, location, date and one or more pre-defined categories (Clough et al., 2005). The IAPR TC-12 Benchmark contains 20,000 still natural images. This collection differs from the St. Andrews collection in two major ways: 1) it contains mainly color photographs and 2) it contains semi-structured captions in English but also in German (Grubinger et al., 2006).

In Chapter 5 (§5.5.2.4), we show how the MAPs obtained for visual retrieval with DORIS in our image set considerably surpasses the best ImageCLEF scores. As previously justified, this difference is made by both our more homogeneous image set and an image retrieval tool (i.e., DORIS) tailored to the features of our prototypical image.

2.2.4. Domain-specific CLIR

The BC model is intended to be used in bilingual specialized contexts. This fact makes domain specific cross-language information retrieval (CLIR) of interest to the present work. While information retrieval (IR) and CLIR have been mostly applied to retrieve information from general collections, some works have been done with the purpose of retrieving cross-lingual information in specific domains following the CLIR methodology. Many of these works have been undertaken in the frame of the CLEF's domain-specific task³⁰. This track has addressed CLIR using the GIRT-4 German/English social science database (over 300,000 documents) and two Russian corpora: Russian Social Science Corpus (RSSC, approx. 95,000 documents) and the ISSS collection of sociology and economics documents (approx. 150,000 documents)³¹.

²⁹ <http://www.st-andrews.ac.uk/imu/imu.php?request=home>

³⁰ <http://www.clef-initiative.eu>

³¹ These figures correspond to CLEF 2006.

As CLIR strongly relies on machine translation, it is widely accepted that machine translation can reach a decent, not *publishable* performance with general texts but not as good for specialized texts –with the exception of some knowledge-based restricted proposals. Likewise, while general purpose lexicons are used for CLIR, similar approaches for domain-specific tasks have to use specialized thesauri or terminological databases. For instance, the thesaurus available for the GIRT data has been used with the purpose of query expansion in the works of Hackl and Mandl (2005) and Petras (2005).

However, in spite of the fact that specialized discourse poses additional challenges for IR, most works have addressed the problem with standard CLIR approaches. In the particular case of the GIRT collection, this trend can be explained by the characteristics of a social science corpus which might be said to be lexically closer to a general collection than, say, an electronics corpus. This is the reason why acceptable results have been achieved with this collection using the same available machine translation software that has also been used for translating documents or queries from general discourse.

We assume, however, that standard CLIR approaches do not perform appropriately in our search space due to the characteristics of the language use described in Chapter 3 (specially §3.4.2). Let us analyze just one example of the possible problems that automated language processing would face. In the following term extracted from a Spanish document, we see recurrent patterns in online catalogs such as shortenings, omission of prepositions, and misuse of punctuation. The complete form of the term is presented on the right after the arrow:

Reten 22+32+5.5 arranque Lambretta → Retenedor 22+32+5.5 de arranque para Lambretta*

While *reten* here is a shortened form of the noun *retenedor*, it is also the third person plural of the verb *retar* (to challenge), or with orthographic accent *retén* means *squad*. Also, as for the syntactic level, the absence of prepositions would lead the automatic translator to interpret *22+32+5.5 arranque Lambretta* as adjectives according to the Spanish syntax, besides other possible problems.

With the BC model, then, we are trying to obtain a trade-off between the recurrent non-standard use of language in online catalogs and the typical clean and homogeneous layout of artifact images described in Chapter 5 (§5.5.2.5). It is with the purpose of addressing these

problems and advantages in our search space that multi-word term (MWT) recognition, noun classification, and image term alignment are discussed in the second part of this chapter below.

2.3. MWT and artifact noun recognition

Our first inquiries about the feasibility of implementing the BC-model shed light on the challenges faced by the model. Besides the image retrieval problem already discussed above, it was necessary to solve non-trivial linguistic issues such as term recognition and extraction, and noun classification. While these problems are justified in the Introduction chapter (§1.2), the methods and techniques used later, mainly in Chapter 4, are grounded here in the subsections below. The review below is framed using representative works to illustrate the latest trends in each track, but also to refer to pioneer contributions that are still valid nowadays.

2.3.1. Multi-word term recognition

After thinking about the possible ways to address the review of the literature related to term recognition, we concluded that our departure point should be the characterization that we have made of our core problem and of its various components. That is, the relevance of the works published so far is determined by 1) their degree of applicability in a search space as the one described in Chapter 3; 2) their adequacy to the term types involved in the BC hypothesis; and 3) their consistency with the purpose and application of the present work.

With these criteria in mind, the methods for term recognition were investigated. A review of the pioneer approaches, specifically the one made by Cabré et al. (2001), suggested two prevailing frames, namely, linguistic and statistical methods. It was interesting to see in a more recent evaluation by Vivaldi and Rodríguez (2007) that these similar trends continue, that hybrid approaches consolidate, and that nearly all approaches draw either upon *termhood* or *unithood*, especially for MWT recognition; unithood defined as the degree of strength or stability of syntagmatic combinations or collocations, and termhood defined as the degree that a linguistic unit is related to or represents domain-specific concepts (Kageura and Umino, 1996).

With the amount of data collected for this thesis (see Chapter 3), statistical methods readily come to mind as an attractive alternative. Statistics for term recognition is mainly implemented in the form of association measures (AMs) (cf. Drouin, 2003), although machine learning and other approaches have also been used (see for instance Nazar, 2011). The task has been

addressed mostly as a multi-word expression (MWE) recognition problem since what is measured here is the unithood of expressions, that is, the stability and the frequency of a multi-word sequence, and this is what AMs specialize in. While there are some few common AMs that have been widely used for MWE recognition (e.g., mutual information, t-test, log-likelihood, etc.), determining which one is the most appropriate one is not a trivial issue.

We examined our corpus considering some of the strengths and weaknesses of a few known AMs. However, such examination could be carried out but only in the light of the observations derived from Chapter 3, the characterization of the search space. A major conclusion was drawn from this inquiry: AMs have proved to be useful on certain data, but drawbacks in our search space such as the non-standard use of language, lack of context, redundancy of data, and corpus size, among others, would undermine AM's performance. This hypothetical failure can be explained by the fact that AMs generally base their statistical formulas on the frequency of expressions. Below, we briefly mention how these drawbacks can affect frequency and other aspects for the application of AMs in our data:

- a) *Non-standard use of language.* It has been observed that in certain online documents, especially online catalogs, MWT's syntax tends to undergo certain transformations. Let us illustrate this case with the following example:

Boot Lid Rubber Buffer BMW → Boot Lid Rubber Buffer for BMW*

Boot Lid Rubber Buffer BMW → BMW Boot Lid Rubber Buffer*

The original MWT candidate is on the left and more acceptable rearranged versions are on the right. The rearranged versions put *BMW* as a premodifier or as part of a prepositional phrase. The problem here is that these transformations are too unpredictable to assume that a given MWT term realizes always in the same way. In other words, the same MWT, but with different realizations, would have different frequencies which affect the performance of AMs. Shortenings of words and variation in the use of capitalization would also have a similar effect.

- b) *Lack of context.* Online catalogs use to present information in the form of bimodal pairs, that is, term and image. This means that there is no context for MWTs in this scenario. For context based measures such as the one proposed by Zhai (1997) this becomes a handicap.

- c) *Redundancy of data*. This issue, brought up by Baeza-Yates and Ribeiro-Neto (1999, p. 368), can seriously affect frequency based measures. In Chapter 3 (§3.3.2.1.2), we showed how redundant our search space is.
- d) *Corpus size and sparseness*. Corpus size is important, among other things, for the estimation of prior probabilities used in further equations. Likewise, the smaller the corpus, the greater the effect of sparseness on such probabilities. We have managed to compile a representative corpus for our observations, but a search in another field or even in the same field but in individual websites could come up with a single webpage as the only corpus. In such a case, the application of AMs would lose relevance.

We checked the problems presented above for AMs using a hybrid approach based on some of the strategies proposed by Vivaldi (2001). First, we used an adaptation of Quiroz's syntactic rules (2008, p. 405) to extract terminological noun phrases (NPs) and then pointwise mutual information was used to determine unithood in the extracted NPs. As we will mention later, the rules behave reasonably well, but the results of the experiment with AMs were not satisfactory. We attribute this low performance to the search space characteristics already presented above.

As for machine learning methods, a recent work has been conducted by Judea (2013) for unsupervised compilation of training sets for automatic terminology acquisition. The training, development, and test sets consist of patents. Discriminant features are extracted and used to automatically label unseen data. Although the main goal of this work is the unsupervised generation of training sets for term acquisition, two classifiers were trained and tuned up to assess the utility of the new training sets for term acquisition on patents. The author reports an *F1* score of 0.784 and 0.789 for both classifiers, which seems to prove the usefulness of the training set. With regard to the selected features, while some of them are applicable to other genres, some others are too specific of patents and therefore not relevant for other type of documents. For example, one of the best features is given by the term candidate followed by a figure reference (e.g., "Figure", "Fig."), which might not be discriminant enough in a different setting. However, even when some of the features make the approach seem somewhat overfitted, the reported evaluation suggests that they appropriately model the distribution of terms in patents.

In an interesting attempt to compare methods, Zhang et al. (2008) evaluate 5 different algorithms for term recognition and then combine them all into a weighted voting approach. All of the evaluated algorithms are statistically- or frequency-based. Two test corpora were used for the evaluation: a specialized corpus (the GENIA corpus³²) and a rather general corpus about animals taken from Wikipedia. The value of this study is mainly given by the comparison of the algorithms and the outcomes of the voting method. After the study, a more objective decision on the appropriate method can be made depending on the nature of the corpus. On the other hand, although this work tries to demonstrate that single word terms are as representative as MWTs, the results show that the different algorithms extracted mostly single word units from the general corpus while the specialized corpus yielded mainly MWTs. This also served to prove that the level of specialization of the corpus determines term recognition performance. In this line, it is worth remarking that the high precision scores of the algorithms for term recognition in the GENIA corpus might be due to the fact that this corpus consists of scientific abstracts whose term distribution certainly differs from other sections of scientific papers. This could also have had an effect in the almost negligible improvement contributed by the voting approach.

With regard to linguistic methods for term recognition, it was found out that having a corpus that maximizes the frequency of MWTs makes linguistic approaches fit better from the beginning. In other words, we agree with Morin and Daille (2010) on the importance of compiling the appropriate data to leverage MWT recognition from the very source, as is also shown by Zhang et al. (2008) (cf. Bonin et al. 2010). Thus, the assumption that the more specialized the corpus, the higher the number of MWTs led us to carefully select our data according to the criteria defined in Chapter 3. The corpus collected according to these criteria enabled the exploration of the most used linguistic methods for term recognition which are implemented for experimentation in Chapter 4. Such methods can be grouped in three broader categories, i.e., syntactic rules (or filters), word lists (or seeds), and boundary-based (or bag of words).

As for syntactic rules, since the origins of term recognition, the usefulness of syntactic patterns was proved by authors such as Bourigault (1992), Justeson and Katz (1995), and Jaquemin et al. (1997). Linguistic rules for MWT recognition can be overgeneralizing, though, for they are

³² <http://www.nactem.ac.uk/genia/>

governed by the grammar of the general language. This means, for example, that certain syntactic patterns regularly frequent for MWTs are also typical of free word associations. This is the reason why authors like Estopà (1999, p. 460) and Quiroz (2008, p. 372) suggest supporting the use of productive linguistic patterns with additional morphological, lexical, and semantic information in order to have a better performance in MWT recognition. Following this line, relevant morphology features such as affixes, stems, and Greek and Latin forms are usually combined with syntactic information in an attempt to boost MWT recognition (Nazar, 2011, Estopà et al., 2000).

Some authors, however, deem all possible terminological patterns difficult to predict. Bourigault (1992), instead of starting with syntactic rules, defines the part-of-speech categories that typically are part of what he calls *terminological noun phrases*. The text is parsed extracting only sequences of such predefined categories and excluding the words whose categories do not belong to this predefined set. The relevant sequences are delimited by boundary markers such as verbs, determiners, and some prepositions which are not included in the extracted sequences. Then, he applies some syntactic rules on the extracted sequences thus deriving candidate MWTs.

On the other hand, at the lexical level, Jacquemin (1997) also used a list of terms to find and expand related terms in a text. These lists, which are also known as seeds, have been also used in more recent works. Baroni and Bernardini (2004) use a list of seed terms to compile a corpus as well as other terms from the web. They follow a bootstrapping method and use the terminology extracted in a first phase to initialize a new term search. Likewise, Nazar et al. (2012) also use a list of seed terms to build a taxonomy of domain-specific terms. These latter works, however, are not purely linguistic, for they combine also statistical methods to reach their ultimate goals. It is also worth saying that the seed approach should not be mistaken for the strategy used in term detection proposals, though. Term detection uses dictionaries to detect terms in a corpus just as they appeared in the dictionary, while seeds are used to detect terms in the corpus and then expand the detected terms to recognize new terms³³.

Seed-based bootstrapping methods, however, have been criticized for being affected by what is known as semantic drift, which denotes the extraction of false positives during the expansion

³³ See, for instance, MERCEDES (Vivaldi, 2003) at <http://brangaene.upf.edu/proves/mercedes/indexNetcedes.htm>

stage. Performance in the bootstrapping process decreases as false positives are used for further iterations. This problem is addressed by Ziering et al. (2013), who used an English-German parallel corpus of European patents to control semantic drift by aligning the corpus at the word level. The approach takes advantage of distinctive morphological features in target language words to make sure they are relevant units and to discard false positives during the bootstrapping phase. The method is claimed to be language independent, although POS tagging and linguistic filters can be used for some languages to improve performance. While the process involves sophistication given by the need of a parallel corpus and the use of machine translation and sentence/word alignment, high accuracy is achieved (0.980 for German and 0.955 for English). Ziering et al. (2013a) also tackle the problem of semantic drift by focusing only on coordination patterns. While most of the term recognition proposals discard all but one of the items of the coordination, the authors take advantage of the hyponymy relation in a coordination to reduce false positives and bootstrap lexicons from technical domains by extracting the whole coordination.

At the semantic level, some works include lexical semantics in their approaches for MWT recognition. For example, Vivaldi (2001) uses a hybrid approach and combine syntactic patterns, statistical measures, and lexical senses from EuroWordNet (Vossen, 1998). Likewise, Vivaldi and Rodríguez (2002) use EWN for medical term extraction. Maynard and Ananiadou (1999), on their part, describe an approach that uses both linguistic and statistical information combining syntax and semantics to identify, rank and disambiguate terms.

The great amount of work on MWT recognition reflects the importance and the complexity of the problem. However, there are important differences between the context of most of these works and ours. Most of the published works rely on neat and homogeneous corpora, and we have just the opposite, that is, noisy data. This noise is in the form of the drawbacks described above. Therefore, the term recognition methods for experimentation in this research were selected bearing in mind the characteristics of our search space and the strengths and weaknesses of the existent approaches.

2.3.2. Artifact noun recognition

Considering that concrete nouns, and more specifically, artifact nouns are a core component of the BC model (see §1.2.2.2.2), there was a clear need to adopt a suitable method for automatic artifact noun recognition. During our explorations, however, we found out that

while there has been a long and active discussion on linguistic, psycholinguistic, and philosophical aspects of concrete and abstract nouns (see, for instance, Craig, 1986; Aikhenval'd, 2000; Altarriba et al., 1999), the problem of automatic noun classification has barely been addressed recently; needless to say that there is much less about artifact noun recognition.

In the line of lexical acquisition, the problem of disambiguating countable and mass nouns is the closest to the problem of classifying concrete and abstract nouns. This is the reason why we found particularly interesting a Bayesian model of inductive learning proposed by Bel et al. (2008). The approach uses local morphological and syntactic features to model a typology of Spanish nouns. Then, a finite state automata is used to evaluate the features of nouns in context. The result of this evaluation are vectors with a number of dimensions equal to the number of defined features. These vectors are used later to obtain the likelihood for the Bayesian model.

On the other hand, the creation of WordNet (Miller, 1995) became a milestone for lexical semantics-based research and applications and, therefore, for some of the first works on automatic noun classification. Later, WordNet inspired the inclusion of languages other than English and of other features and this is how EuroWordNet (Vossen, 1998) was born as an independent system for European languages.

Our first inquiries on artifact recognition suggested that the only tool for noun classification at the time was the SuperSense Tagger (SST, Ciaramita and Altun, 2006), which is WordNet-based. The SST, currently available for English and Italian, is a Hidden Markov Model (HMM)-based tagger which uses a probabilistic model to determine hidden values (classes) from observations (nouns and verbs). In practice, this means that the SST does not assign the most frequent sense of a word as other applications do. Instead, it calculates the probabilities by means of a HMM to assign the most appropriate sense of a word according to its context. The senses assigned in the output correspond to EWN's subclasses of origin (*natural* and *artifact*), form, composition, and function.

Later, a similar approach for Spanish but with a different algorithm called UKB was proposed by (Agirre and Soroa, 2009). The authors apply the "so-called Personalized PageRank on a Lexical Knowledge Base (LKB) to rank the vertices of the LKB and thus perform word sense

disambiguation". Contrary to the SST, UKB's output is not the *origin* or the *class* according to EWN, but the most probable sense in the form of a numeric code –as it is in EWN's database files. In addition to this, the UKB can also be set to yield the most frequent sense instead of the most probable sense.

These first approaches to word sense disambiguation suggest that tackling the problem of artifact recognition requires semantic information to improve performance. This explains why some other approaches keep on the same line, although trying to go a little beyond lexical semantics. In 2008, interesting proposals were presented in a dedicated shared task in the lexical semantics workshop in the frame of the European Summer School in Logic, Language and Information (ESSLLI). It seems that the focus for noun classification turns now on distributional semantics. Thus, Katrenko and Adriaans (2008) use Pustejovsky's (2001) description of qualia structures to extract formal, constitutive, telic, and agentive features in the form of syntactic patterns from contexts that include nouns to be classified. Their clustering technique seems to behave well, especially with the *formal* role. Artifacts are acceptably classified, although other categories such as *vegetables* and *animals* are better represented by certain features and, therefore, better ranked.

Likewise, Bullinaria (2008) uses distributional semantics to cluster words of the same class when they appear in semantically related contexts. He found that the set-up producing the best results involved using Positive Pointwise Mutual Information (PPMI), small window sizes (just one context word each side of the target word), and the standard Cosine distance measure. Barbu (2008), on his part, combines association measures with linguistic patterns in order to derive information about hierarchical relations, holonymy and meronymy, location, actions, and inherent properties (cf. Katrenko and Adriaans, 2008).

The results reported by these and other recent research on noun classification are appealing. The concrete/abstract discrimination experiments published by Peirsman et al. (2008) and Van de Cruys (2008), for example, reach an extremely high precision (entropy = 0, purity = 1). These facts compels the community to consider their methodologies to be tested with different datasets.

For the purpose of this thesis, even when the works just mentioned in the previous paragraph seem highly promising, we decided to test artifact noun discrimination in our data following

Bel et al. (2008), Ciaramita and Altun (2006), and Agirre and Soroa (2009). This decision is justified, on the one hand, by the fact that an inductive Bayesian model takes advantage of local clues which makes it less dependent of external resources and at the same time optimizes computational efficiency. On the other hand, the WordNet and EWN-based approaches proved to perform reasonably well in previous experiments (see Burgos, 2009) and are available and established for English and Spanish. It is also worth saying that preliminary observations in EWN of nouns extracted from our data confirmed the adequacy of this resource for this specific task. These preliminary observations in EWN can be summarized as follows: a) Out of 570 nouns for Spanish, 254 nouns have at least one artifact sense; b) out of 1030 nouns for English, 598 nouns have at least one artifact sense; c) when there is only one sense and it is *artifact*, the chance of error is minimum; d) when there is more than one sense, but the first or second sense is *artifact*, the chance of error is minimum too.

The experiments with these approaches, as well as another proposal using non-linguistics variables (Burgos and Wanner, 2006), are presented in detail in Chapter 4. Evaluation of performance of each method is also provided in the same chapter.

3. SEARCH SPACE CHARACTERIZATION

As it was explained in Chapter 1 (§1.5.1.), the BC hypothesis-based MWT retrieval is to be implemented on digital environments, for example, on the Web. However, the amount of information in the Web is so huge and heterogeneous in its structure that it results impractical to think of the whole Web as the search space for the application of the present research's model. Likewise, the bimodal nature of our model implies an image component which certainly is more costly to analyze and to index than raw text. This means, therefore, that an appropriate segment of the Web has to be demarcated.

Such a Web segment is to comply with some basic criteria:

1. *The features of the studied phenomenon.* The Web segment must feature the two components of the bimodal co-occurrence, i.e., images and text, in a representative number of documents.
2. *Accessibility of data.* Current available tools for search space exploration consist of text analysis tools and basic image matching tools. It means, on the one hand, that a certain amount of text in the segment's documents must be accessible for text analysis tools. It then excludes text embedded in images or in flash movies, for example. On the other hand, image matching will be carried out at a basic texture level on prototypical images which are described in Chapter 4. Thus, a representative number of images are expected to have the features of the prototypical image to guarantee a minimum of adequacy of the Web segment in this sense. In short, the characteristics of text and images in the selected segment must enable the available tools for their analysis.
3. *Authority.* The authority of documents constitutes an important criterion for it affects quality minimums such as: a) usage of standard terminology; b) appropriate use of markup languages such as HTML and XML; c) best programming practices which result in the use of informative meta-tags, well-structured sites, etc.; and d) homogeneity in image and text characteristics and quality.

3.1. Outline

According to the above considerations, it was observed that a suitable search space for this thesis could be a manually controlled and categorized Web segment, that is, a Web directory. Baeza-Yates and Ribeiro-Neto (1999) define Web directories as hierarchical taxonomies that classify human knowledge, although they also note that these taxonomies have cross references which really makes them directed acyclic graphs.

Web directories have motivated a lot of research. Their value as a semantically labeled Web segment has been remarked and exploited by different works (see, for instance, Perugini, 2008, Osiński and Weiss, 2004, or Santamaría et al., 2003). Web directories have been used as research objects in themselves, but also as the basis for the study, experimentation and evaluation of a wide range of phenomena and methods³⁴. It is in this sense that a particular Web directory is used in the present research. We take advantage of its contents and structure to test and apply our model to MWT retrieval.

The Web directory chosen as the Web segment for our research is the **Open Directory Project** (ODP)³⁵. For an initial description of what the ODP is, let us use and adapt the ODP's own description^{36,37}:

The Open Directory Project is the largest, most comprehensive human-edited directory of the Web. It is constructed and maintained by a vast, global community of volunteer editors. It is hosted and administered by Netscape Communication Corporation. Netscape administers it as a non-commercial entity, and claims to be committed to keeping it a free and open resource via its social contract with the Web community.

The ODP is a Web directory, not a search engine. Although it offers a search query, the purpose of the ODP is to list and categorize Web sites; not to rank, promote or optimize sites for search engines. The ODP is simply a data provider. The ODP powers core directory services for some of the most popular portals and search engines on the Web, including AOL Search, Netscape Search, Google, Lycos, HotBot and hundreds of others.

³⁴ For a descriptions of how web directories are structured, see, for instance, Baeza-Yates and Ribeiro-Neto (1999:384) or Perugini (2008).

³⁵ <http://www.dmoz.org/> retrieved April 6th, 2011

³⁶ <http://www.dmoz.org/about.html> retrieved April 6th, 2011

³⁷ <http://www.dmoz.org/help/geninfo.html> retrieved April 6th, 2011

The description above initially responds to some of the requirements previously established for a suitable Web segment. As we will show later, the ODP includes a representative number of accessible images and text in different languages. Likewise, compared to the wide range of documents uncontrollably published in the whole Web, the ODP constitutes an authoritative source since it is constructed and maintained by human editors and administered by a recognized international entity (cf. Santamaría et al., 2003).

The sections below describe the characteristics of our Web segment and, at the same time, justify its selection as search space for the present research. First, the macrostructure of data is discussed and some general figures are presented; a further delimitation of the Web segment has to be done due to the large number of categories included in the ODP; and the distribution of categories per language in this new subsegment is presented. Second, an analysis of the data microstructure is carried out. Here, we studied the strings in the uniform resource locators (URLs) of our search space, keywords and descriptions in the HEAD section of any site in our search space, and the contents of the BODY section, which are mined to observe representativeness of nominal MWTs and, particularly, artifact MWTs. Finally, the crawling methodology as well as some drawbacks of the search space is presented.

3.2. Data macrostructure

As for the data, the ODP provides open access to the structure and contents of the whole directory. Access to this information is provided by means of dump files in XML format³⁸. The structure file is 860 MB and has 11,578,462 lines. It contains all the paths in the different categories from the root to the leaves. The general file structure is as follows:

³⁸ <http://www.dmoz.org/rdf.html> retrieved on March, 2011

Field	Field value
<Topic>	Category name (tag opens)
<catid> </catid>	Category ID
<d:Title> </d:Title>	Category title
<lastUpdate> </lastUpdate>	File update date
<d:Description> </d:Description>	Category description
<narrow> </narrow>	Category's children
<narrow1> </narrow1>	Category's children
<narrow2> </narrow2>	Category's children
<altlang> </altlang>	Category in another language
<related> </related>	Related categories
<symbolic> </symbolic>	Cross reference
<symbolic1> </symbolic1>	Cross reference
<symbolic2> </symbolic2>	Cross reference
<Alias>	Alternative name for a category (tag opens)
<Target> <Target/>	Target category for the Alias
</Alias>	Alias (tag closes)
</Topic>	Category name (tag closes)

Table 6. Fields of ODP structure file

The content file is 1.9 GB and has 28,877,722 lines. It contains all the paths and the URL of every single site in every category from the root to the leaves. The content file structure is as follows:

Field	Field value
<topic>	Category name (tag opens)
<catid> </catid>	Category ID
<link> </link>	URL
</Topic>	Category name (tag closes)
<ExternalPage>	URL (tag opens)
<d:Title> </d:Title>	Title of site
<d:Description> </d:Description>	Description of site
</ExternalPage>	URL (tag closes)

Table 7. Fields of ODP content file

The structured information in the content and structure files allowed for systematic study and extraction of relevant paths and sites. Likewise, it makes possible a characterization of categories per language, as described below.

3.2.1. General figures

As a directory, the ODP clearly differentiates itself from the rest of the Web because it presents a hierarchical structure which in turn is thematically ordered. A set of websites belonging to the same category are grouped under a common node of the structure even when they are in different languages. Thus, instead of finding relevant websites using keywords in a search engine, this hierarchical structure enables the user to select a category and to follow a path until the desired category is reached or until no more categories but websites are found. In this context, a path is a track that can be followed from a root node to a leaf node. For instance, the paths below can be used to find websites in different languages related to the automotive industry:

- ◆ Negocios : Industrias: Automotriz (spa³⁹)
- ◆ Business : Automotive (eng)
- ◆ Affari : Veicoli : Automobili (ita)
- ◆ ビジネス (Business) : 自動車 (Automotive) (jap)

Up to March, 2011, the ODP reports $\approx 4,868,292$ sites classified in $\approx 1,006,417$ categories; a work done by 90,614 human editors. It has been continuously growing up since it was born in 1998. Different stages of this growth are described by Osiński and Weiss (2004) who reported 575,000 categories and by Perugini (2008) who informed about 689,000 topics or categories.

However, the reported figures by DMOZ on the number of categories and sites per category should be revised or at least clarified. As for the number of categories, when the DMOZ site reports $\approx 1,006,417$ categories, it sounds as if there were such a huge number of unique categories. But once the file containing the ODP structure is analyzed, it can be found out that in fact there are 318,340 unique categories in different languages⁴⁰.

³⁹ ISO language codes

⁴⁰ To draw these figures and compare them with the DMOZ's ones, all the unique topic paths were extracted from the ODP's structure file. Then all the categories were sorted and duplicates were removed.

As it is not clear enough from DMOZ site, it could be hypothesized that when they report such a big number of categories, it is because:

- a) DMOZ's numbers refer to paths rather than to categories. The structure file contains 771,007 unique topic paths, though. We assume that summing up newsgroups and some other non-topic categories, the number of paths could reach the figures given by the DMOZ site.
- b) DMOZ numbers suggest that a category x may belong to different paths and that in some of its occurrences it could be a child of different parents. The number of categories would increase just in case different occurrences of x do not refer to the same category, but are homonyms of each other. If it is true, the number of categories would increase in a certain degree but not so much.

In the line of the latter possibility, it is true that there is not always just one path to a given category. For example, for English we found that *Parts_and_Accessories* is a child of different parent categories. However, even though the category *Parts_and_Accessories* has different parents in many cases, there is no semantic change in the category. For example, it is possible to reach the *Parts_and_Accessories* category by following either of the paths below:

Business : Automotive : Motorcycles : Parts_and_Accessories

Shopping : Vehicles : Watercraft : Parts_and_Accessories

Regional : Europe : United_Kingdom : England : Worcestershire : Business_and_Economy : Motoring : Parts_and_Accessories

Thus, at least for the *Parts_and_Accessories* category, which is of particular interest for the present research, the fact of having different parents does not affect the semantic relation which groups the occurrences of the category under the domain of automotive engineering. The section below explains how the analysis of *Parts_and_Accessories* category was performed and delimited.

3.2.2. Category delimitation

Out of the total of categories of the ODP, just those categories belonging to the subject-field relevant to this thesis, i.e., automotive engineering, were kept. It was necessary to take into account the depth of paths under the category *Automotive* in each language in order to determine the subcategory or subcategories whose websites would be defined as part of our search space.

For example, for English, *Automotive* is a non-leaf category, that is, under *Automotive* the user can find some web sites but there are also further subcategories to keep browsing. On the contrary, for Spanish, *Automotriz* is a leaf in most of the paths, i.e., it does not have subcategories but a set of sites classified in under it:

- ◆ (spa) Negocios : Industrias : **Automotriz** : www.site1.com; www...
- ◆ (eng) Business : **Automotive** : Parts_and_Accessories : Electrical : www.site1.com; www...

Shorter paths for Spanish, however, do not mean that sites under the Spanish category *Automotriz* cannot be additionally subcategorized. There are sites in this category that could perfectly be classified in a category called, say, *Partes_y_accesorios*. And, in fact, there is a category called *Autopartes*, but it is only as a leaf in the path *Regional : América : México : Economía_y_negocios : Industrias : Automotriz : Autopartes* and not directly in *Negocios : Industrias : Automotriz*.

Should there be a finer classification in Spanish, it would be possible to find categories at a lower level matching with same-level categories in English or in other languages. If both languages coincided deep at a low level category, it should suffice to assure comparable corpora in a particular language pair, and to reduce the search space for faster and more effective uses of translation or corpus linguistics techniques. As an illustration of an ideal scenario of two languages with similar depths in their paths for a specific category, let us consider the English-Italian language pair. For both languages, there is a common path to the category *Electrical* for English and *Elettrici_ed_Elettronici* for Italian:

Business : Automotive : Parts_and_Accessories : Electrical
Affari : Veicoli : Automobili : Parti_e_Accessori : Elettrici_ed_Elettronici

For English, there are 67 sites and for Italian there are 21 websites under the leaf category. Both sets of sites should constitute an interesting bilingual comparable corpus in the subdomain of automotive electrical parts and which can be reached by following common paths of the ODP. A word frequency analysis of keywords and descriptions of sites for English

and Italian under these categories, confirms the close thematic relation between both corpora (see Figures 12 and 13):

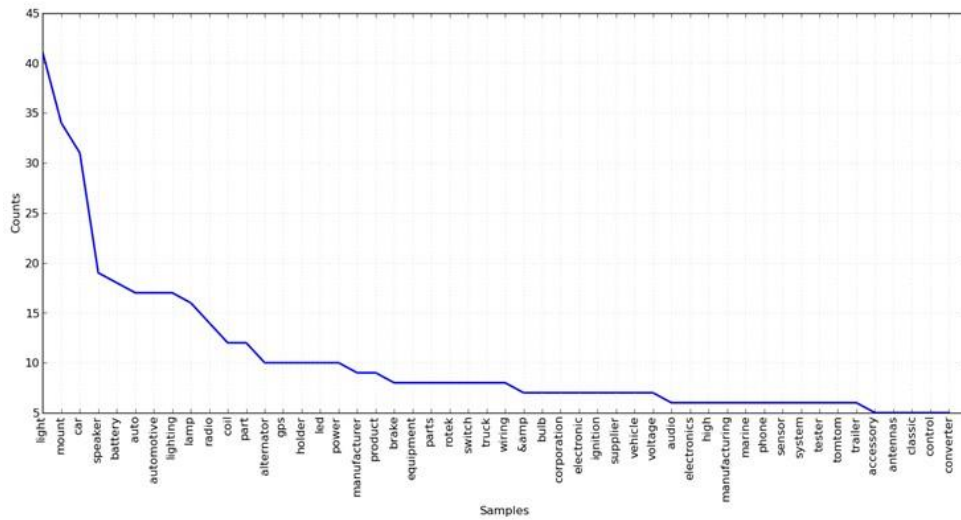


Figure 12. Lemma frequency analysis of keywords and descriptions for English

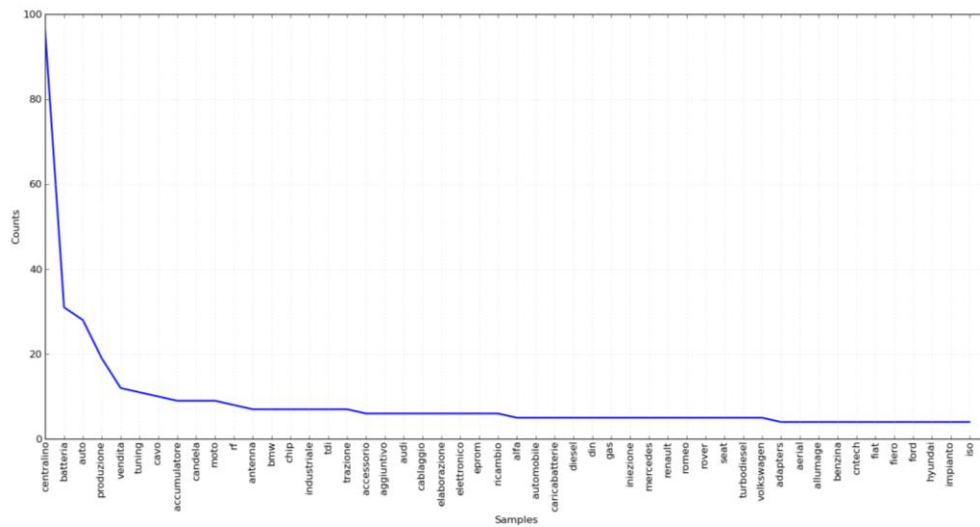


Figure 13. Lemma frequency analysis of keywords and descriptions for Italian

From the analysis, it can be observed that there are similarities in the distribution of some equivalent words in the English-Italian language pair.

battery: *batteria* or *accumulatore*

car: *auto*

wiring: *cablaggio*
accessory: *accessorio*
electronic: *elettronico*
part: *ricambio*

This distribution preliminarily suggests that this category in both languages groups documents which constitute a potential comparable corpus.

Unfortunately, there is not the same granularity in the ODP for Spanish classification to match corpora with the same degree of precision with other languages like English. This is an issue which certainly poses additional challenges. For example, such a high level category as *Automotive* groups sites ranging from part manufacturers to news and media, transport, associations, etc. This means that the language with a finer classification will yield less noise form non-relevant categories than a language with shorter paths or subcategorization, which is the case of Spanish.

Notwithstanding this lack of ideal subcategorization for Spanish, the available data and categories is what we have as starting point and, therefore, the sites under the *Automotriz* category constitutes the Spanish search space for our model. On the English side, however, some further decisions can be made. Given that there are so many sites and subcategories under the whole English category *Automotive* compared to the Spanish search space, just sites under one English category were included, namely, the *Parts_and_Accessories* category.

The section below shows how the structure file was filtered to extract just the relevant instances of the respective categories for both languages.

3.2.3. Category filtering and distribution

For English, as said before, we found that *Parts and Accessories* is also a subcategory of other main categories besides *Business* (e.g., *Home*, *Recreation*, *Regional* and *Shopping*). Therefore we filtered by *Business*, *Regional* and *Shopping* and discarded non-relevant subcategories. Thus, we found out that in the following level, besides *Automotive*, there were other subcategories such as *Europe*, *Music*, *Oceania*, *North America*, *Sports* and *Vehicles*; we excluded *Music* and *Sports*. Once applied these filters, 8 subcategories remained at the level of parts and accessories: *Aircraft*, *Autos*, *Motorcycles*, *New Zealand*, *Parts and Accessories*, *United Kingdom*, *United States* and *Watercraft*. As all of them could potentially embrace relevant sites for this research, they were kept. The

- *Shopping*: contains sites of which the primary focus is to allow the consumer to select and obtain goods and services over the Web⁴¹.
- *Business*: lists and categorizes English-language sites that cover business as an activity and business as an entity⁴².
- *Regional*: contains English language sites about geographical regions. This includes groups of countries in an area, individual nations, states or provinces and localities. The top Regional category covers sites that are global in scope⁴³.

According to these descriptions, and considering that most of the categories depend on *Shopping*, it could be inferred that there should be a representative number of product catalogs. It should be so in order to enable users to select and order the desired products. The description of *Business* also suggests the presence of visual and textual information since Business as an activity includes “official Web sites for and about corporations and commercial enterprises (including subsidiaries) that manufacture, distribute, market and sell goods and services to other businesses (B2B) and/or consumers (B2C)”. As for the *Regional* category, the description suggests an interesting source or repository of terminological variants. This can be concluded from the fact that sites listed under this category produce, sell or distribute products for a delimited geographical region.

For the Spanish search space, there are two main categories: *Negocios* and *Regional*. In the next level, there are the categories *América*, *Europa* and *Industrias*. Then, there are categories corresponding to Spain, Central and South America countries as well as the category *Automotriz*. The next level has the categories *Economía_y_Negocios* and *Comunidades_Autónomas*. Then, there are categories corresponding to some regions of Spain as well as the category *Industrias*. In the next level there are categories corresponding to some subregions of Spain as well as the categories *Economía_y_Negocios* and *Automotriz*. The next level includes the categories *Autopartes*, *Carrocerías*, *Economía_y_Negocios*, *Fabricantes*, *Filtros*, *Industrias*, *Lubricantes_aceites_y_grasas*, *Motocicletas*, *Motores*, *Neumáticos*, *Reductores_y_Engranes*. The category

⁴¹ <http://www.dmoz.org/desc/Shopping>

⁴² <http://www.dmoz.org/desc/Business>

⁴³ <http://www.dmoz.org/desc/Regional>

Lubricantes, aceites y grasas was excluded. In the next level there are the categories *Automotriz* and *Industrias*. And, finally, there is the category *Automotriz*.

For Spanish, 34 relevant paths, out of 35, remained including the category *Automotriz*. Figure 15, illustrates the Spanish search space after the filtering process where just the relevant categories were kept. It can be observed that most of the relevant categories (38) depend on *Regional* while *Negocios* account for just 2 categories. After the Figure, it is also clear that *Automotriz* is in the intersection of the two root nodes along with *Industrias*.

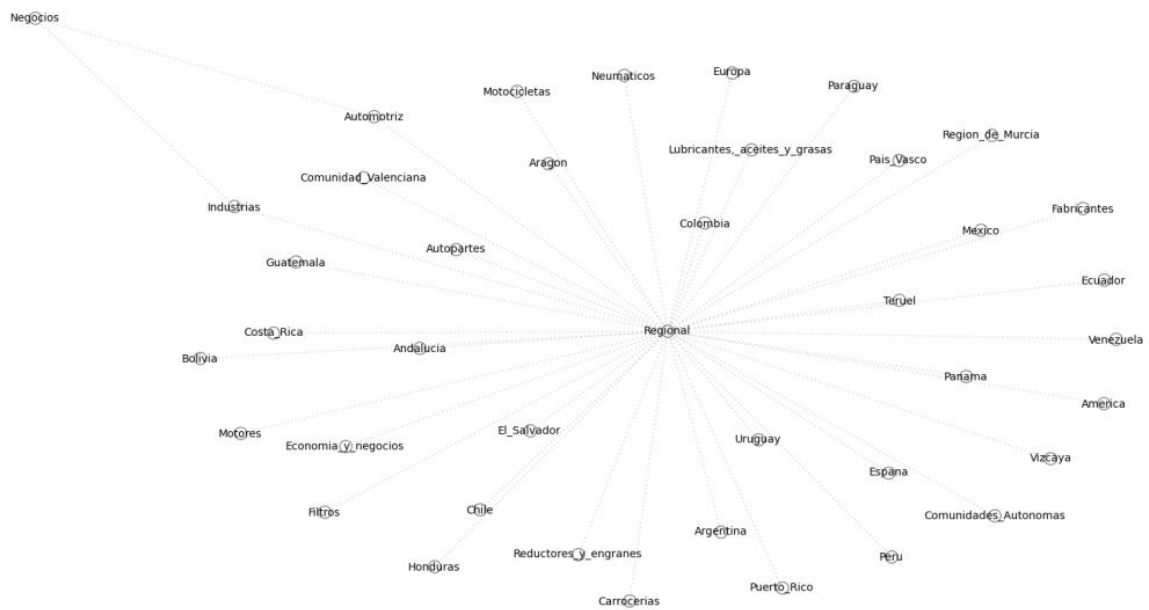


Figure 15. Distribution of Spanish search space

In order to give further interpretation of the Figure, let us review the DMOZ's descriptions for each of these main categories in Spanish:

- *Regional*: contains links to sites with activities in specific countries whose contents are in Spanish.
- *Negocios*: contains links to sites in Spanish of international scope. It lists sites according to existent classifications in order to present a section fully oriented to business activities where users find information on products and companies from various industrial, commercial and service sectors.

Given that most of the categories depend on *Regional*, it is expected that sites under this category represent an interesting source of terminological variants in Spanish. It is also more probable that sites under the *Automotriz* category respond to a regional scope, and that there will be fewer sites with international scope which are those under *Negocios*.

The above-described configuration affects the visibility of and access to sites in the search space. For example, for English the path *Business : Automotive : Parts_and_Accessories* leads to 953 sites⁴⁴. However, as it was showed, once all the occurrences of *Parts_and_Accessories* in the structure file are extracted and filtered, it can be found out that there are other 150 relevant paths for *Parts_and_Accessories*. The analysis of these 151 paths in the content file leads to $\approx 2,177$ sites of the category.

Likewise, for Spanish, the path *Negocios : Industrias : Automotriz* leads to 63 sites⁴⁵. However, once all the occurrences of *Automotriz* in the structure file are extracted and filtered, it can be found out that there are other 34 relevant paths for *Automotriz*. The analysis of these 35 paths in the content file leads to ≈ 320 sites of the category.

Unfortunately, although all these instances of the *Automotriz* and *Parts_and_Accessories* categories belong to the same domain, i.e., automotive engineering, sometimes there are no clear links between them, even within the same language. For instance, the *Parts_and_Accessories* category under *Business*, *Regional* and *Shopping* are not clearly connected. It could cause that once the intuition has led the user to one instance of the category under, say, *Business*, the other instances will not probably be noticed by the user because they are not visible enough from the current instance of the category.

This problem of unclear links or cross references between related categories has to do with the acyclic nature of the ODP. These links, presented in the ODP structure as symbolic links, were studied by Perugini (2008). His study shows the acyclic structure of the OPD and derives some interesting considerations which can be useful to explain the lack of evident interconnections between related categories:

⁴⁴ <http://www.dmoz.org/Business/Automotive/> retrieved April 6th, 2011

⁴⁵ <http://www.dmoz.org/Business/Automotive/> retrieved April 6th, 2011

- Nearly all (>97%) of the symbolic links in ODP create multiclassification,
- Most (>89%) of those multiclassification links connect topics within the same top-level category of the root rather than bridging two distinct top-level categories,
- While the fraction of total multiclassification links that connect two distinct top-level categories is very small (<11%), those links cover over 77% of the possible, distinct top-level category–category connections,
- While the fraction of total multiclassification links that connect two distinct topics (on different hard paths) within the same top-level category is very large (>89%), only a small percentage (10%) of those connect two distinct immediate sub-categories of the same top-level category (<9% of all symbolic links), and
- The majority of symbolic links (>77%) are multiclassification links which connect two categories which share at least the first two levels of topic specificity.

These findings suggest that this certainly is not a trivial problem where the user needs to know a priori all the instances where the relevant categories are located, which is improbable; alternatively, the directory search query service can be used by the user to reach the required category or website, but it is an option which seems not to be optimized or prioritized by DMOZ, as stated in §3.1 above.

3.3. Data microstructure

The previous sections sketched the analysis, filtering and delimitation of the search space macrostructure. Likewise, the ODP categories and subcategories relevant for this research were defined and described. Now, with the purpose of characterizing the search space at a finer level, two major blocks of information have been studied: a) Strings contained in URLs of websites under the relevant categories, and b) the text content in HEAD and BODY sections of websites under relevant categories. These analyses shed light on the URLs that should be emphasized to optimize search space crawling. They also provided us with clues on the type of information and the number and quality of MWTs and images that will be found in the search space.

3.3.1. URL analysis

In order to optimize search space crawling, a sample of the URLs pointing to webpages under relevant categories of the ODP was extracted. The strings in the URLs were analyzed to determine those URLs potentially containing instances of the bimodal co-occurrence. The assumption here is that URLs containing strings like *catalog*, *product*, *part*, etc. maximize the probability of leading to instances of the BC since webpages pointed by these strings tend to include images with their indexes or descriptors. For example, the URL http://www.nosso.com.ar/spanish/catalog/results_search01detail.php?CodProd=ZEN%200404 contains strings like *catalog*, *search* and *prod* and leads to a web page with an instance of the BC:

A string in this context is defined as any sequence of alphabetical characters. The string separator, therefore, is any other non-alphabetical character. For string counting, any single URL in the sample was put in a new line of a file. The file was tokenized using separators. Tokenized files were lowercased⁴⁶ and strings were counted.



Figure 16. Instance of the BC Hypothesis in a web catalog

The total of strings for the URL sample in the English search space is 8,536,423 and for Spanish the total of strings is 762,127. Figures 17 and 18, show the most frequent strings in URLs for both languages.

⁴⁶ Strings were not lemmatized since some strings are intentionally shortened in URLs, so that lemmatization does not make sense in many cases)

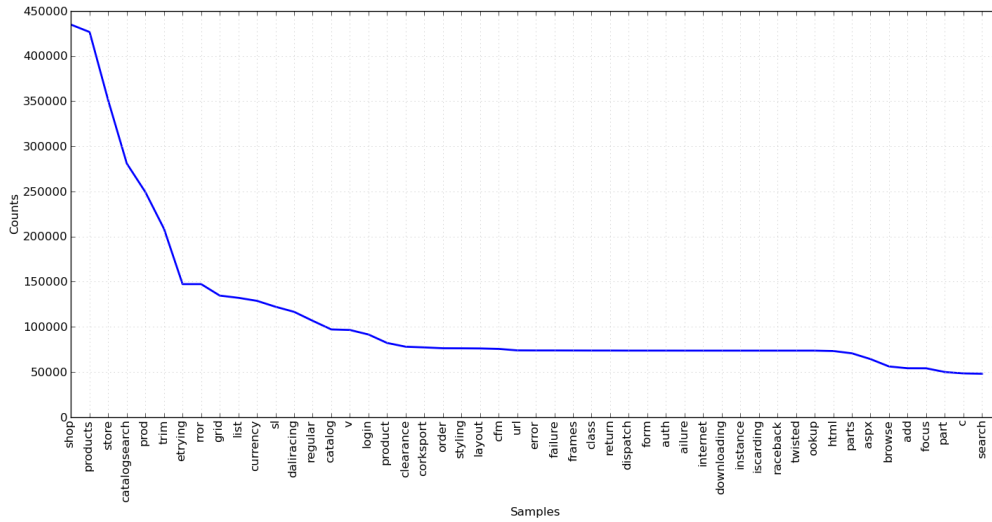


Figure 17. Most frequent strings in URLs for English search space.

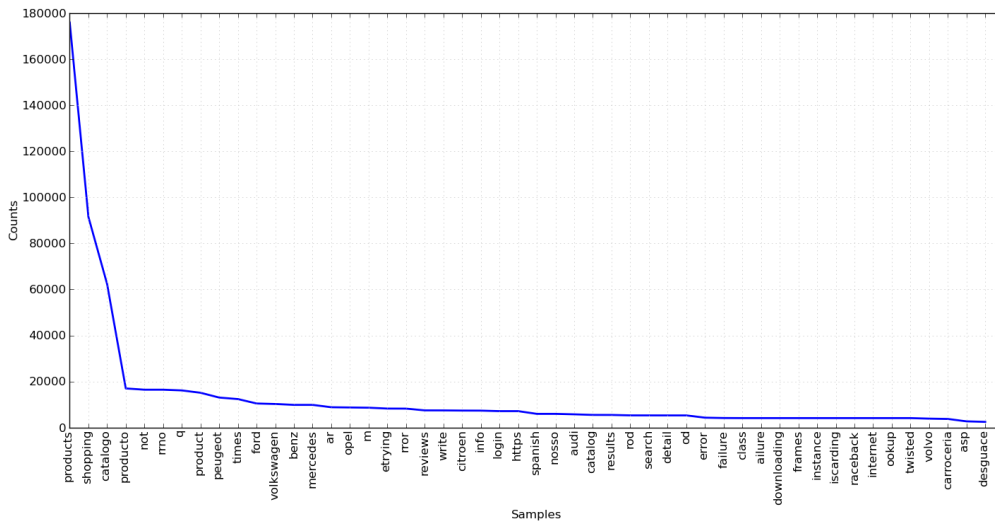


Figure 18. Most frequent strings in URLs for Spanish search space

From this analysis, URLs potentially containing instances of the BC can be identified. Strings like *shop*, *products*, *store*, *catalogsearch*, *prod*, *catalog*, *product* and *part* for English, and strings like *catalogo*, *producto* and *shopping* for Spanish are frequent and seem to be characteristic of URLs containing instances the BC. There are also some common productive strings for both languages like *product(s)* and *catalog* (see Table 8).

String	English count	Spanish count
Shop	434822	72
Products	426624	175892
Store	350850	1
catalogsearch	280927	0
Prod	249242	0
Catalog	97106	5503
Product	82143	15121
Part	49935	1
Stock	899	23
parte(s)	3	0
Catalogo	0	62350
Product	0	17001
Tienda	0	3
autoparte(s)	0	221
recambio(s)	0	888
Productos	0	2469
Shopping	561	91492

Table 8. Most frequent strings for URLs in English and Spanish search space.

Once the most frequent strings potentially pointing instances of the BC were identified, webpages with URLs containing those strings were crawled. These selected webpages are object of observation to analyze number and characteristics of the BC components, that is, text and images. This is precisely the analysis described in the sections below.

3.3.2. Analysis of BC components

3.3.2.1. Text analysis

In order to characterize the linguistic component of our search space, an analysis of two sections of the global structure of HTML documents was carried out, namely, the HEAD section and the BODY section. From the HEAD section, we analyzed the meta data of Description and Keywords tags. The analysis of these meta tags shows frequency, distribution and specificity of terms. As a whole, they provide an overview of the relevance of the information that was going to be found in the BODY section.

From the BODY section, the text was processed and analyzed. The analysis shows token/type ratios as well as relative frequencies of nominal and artifact MWTs in the search space. The results here reflect the degree of relevance of the information in both languages as for the representativeness of artifact MWTs in the search space.

3.3.2.1.1. The HEAD section

For the HEAD section, a word frequency count was carried out. It must be noted, however, that not all the sites contain values for the Keyword and Description meta tags; 162 Spanish sites (out of 320) and 1,665 English sites (out of 2,177) had keywords. Likewise, 161 Spanish sites and 1,643 English sites had a description. Besides, even though the DMOZ's instructions for editors are clear not to include Spanish websites with descriptions in other languages, some of them do have descriptions and/or keywords in English.

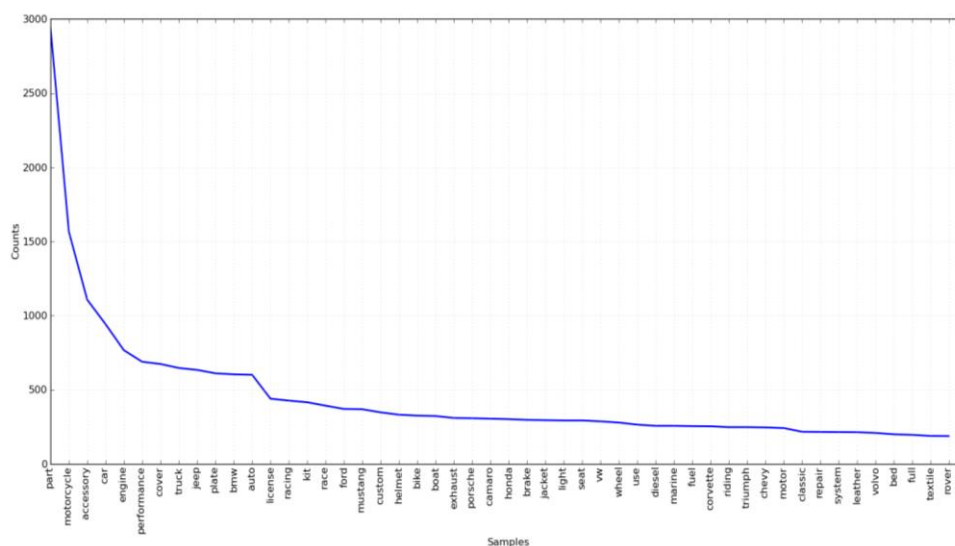


Figure 19. Lemma distribution of keywords in English search space.

After the word frequency analysis, the English search space seems homogeneous and coherent with the categories of the ODP. Figures 20 and 21 show the word frequency for the 50 most frequent words in the keywords and description meta tag values of English sites. It can be seen that the four most frequent words are the same in both Description and Keyword meta tags (*part*, *accessory*, *motorcycle* and *car*). On the other hand, it is interesting to note that the two most frequent words coincide with the name of the OPD category selected for this study, that is, *Parts and Accessories*. It is interesting because websites in this category seem to be developed with

independence of the ODP and were added later to the *Parts and Accessories* category. Likewise, the distribution of frequent words in both keywords and description behaves very similarly too. For the analysis, the text was lemmatized and function words were excluded.

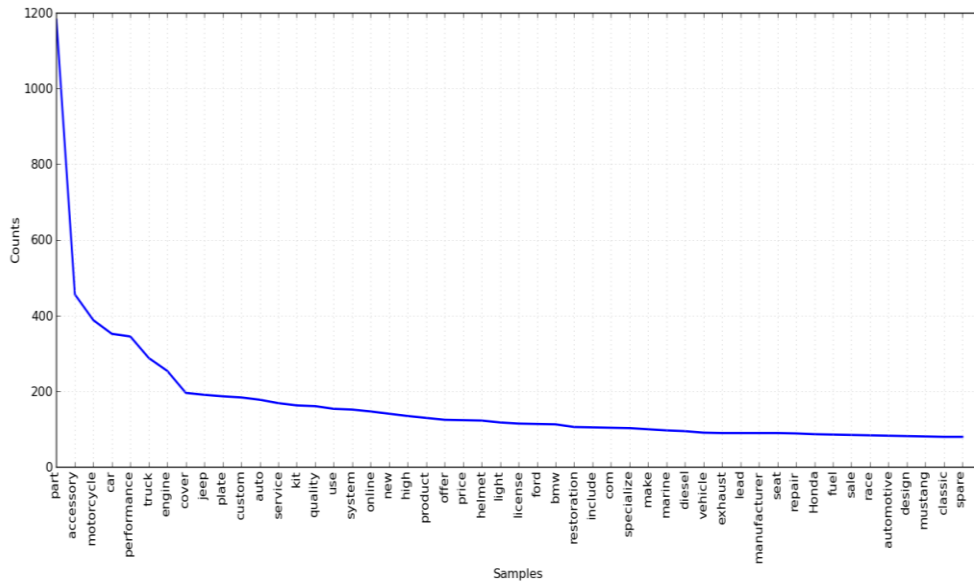


Figure 20. Lemma distribution of descriptions in English search space.

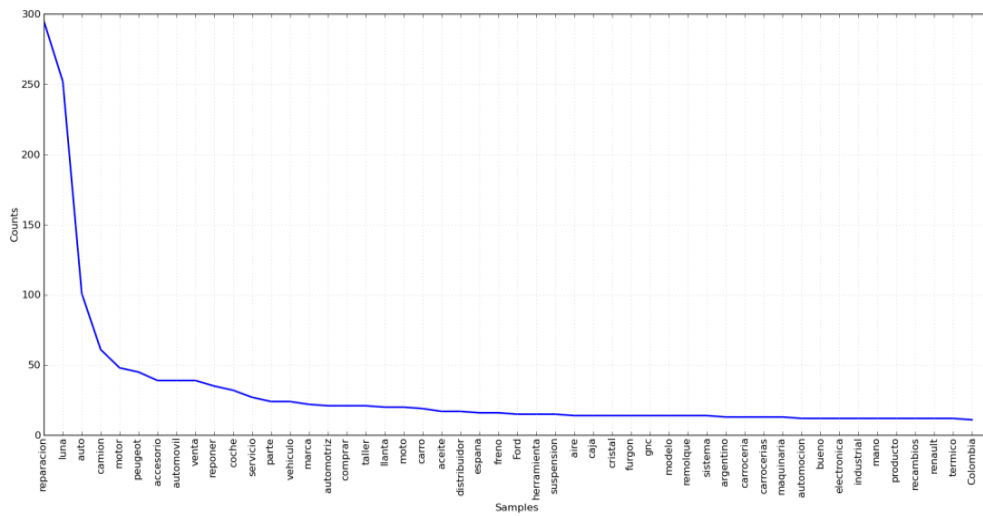


Figure 21. Lemma distribution of keywords in Spanish search space.

For Spanish, however, there seems to be higher variability (see Figure 22). There appears to be a direct relation between the broader scope of the category selected for Spanish, i.e., *Automotriz*, and the occurrence of more general domain words (e.g., *reparación*, *venta*, *servicio*) in the analysis. The fact that *Automotriz* is a high level category could explain the wide semantic range of terms. It is worth noting, however, that terms somewhat more relevant for this study like *motor*, *accesorio*, *parte* and *producto* are reasonably well ranked. For the analysis, the text was lemmatized and function words were excluded.

The difference between English and Spanish is evident in terms of the informational capacity of the selected meta tags. The English analysis lets foresee a well-delimited and more specific search space as for the interest of this research. The Spanish analysis suggests that the search space is still relevant although a higher occurrence of non-relevant MWTs is expected.

The general results of this analysis are, however, as expected. It makes sense to have more general terms in the Description and Keywords meta tags, considering their function. These meta tags are not the natural place for highly specialized MWTs to appear; these must occur in the BODY section of the websites where, in turn, the frequency of more general terms should decrease.

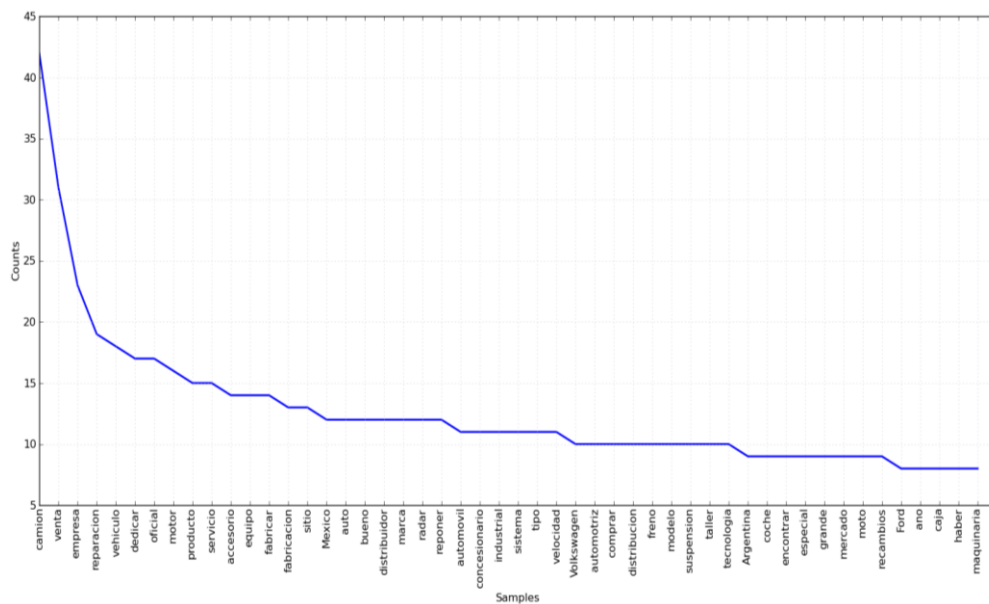


Figure 22. Lemma distribution of descriptions in Spanish search space.

3.3.2.1.2. The BODY section

For the text analysis of the BODY section, 30 sites in English and 36 sites in Spanish were crawled⁴⁷. The methodology below shows how these sites were processed and their text analyzed after crawling⁴⁸:

a. Remove and replace non-text characters:

- 1) *Remove non-text from PDFs.* The spider used for crawling also grabbed some code corresponding to PDF documents which generated a good deal of strange characters. Such code was removed since it hinders further processing and, besides, PDF content is not within the scope of this research.
- 2) *Replace squares by "ó".* It was difficult to predict how every single site was codified. After crawling, it was found that some accented characters, mainly in the Spanish sites, had been printed in files as non-text characters (e.g., as little squares or triangles). In order to prevent processing problems at a later stage, every non-text character was replaced by an accented *o* (ó). This decision was made after a Spanish dictionary-based analysis which showed that out of the five vowels and the *ñ*, the letter that is accented most often is the *o* (35.4%), followed by the *i* (26.8%). The extracted MWTs at the end of the process will be checked and fixed when necessary with the right vowel.

- b. **Normalize lists.** As a general rule, terms or phrases in a list do not end with a period, as in the following example:

Cooling System
Electrical Components
Engine Electrical Components
Engine Filters

After tokenization, the tagger would read these four terms as one single sentence because they do not have a period after the last word to mark the sentence boundary.

⁴⁷ Technical and time constraints did not allow to obtain the text information from all the sites in our search space

⁴⁸ The methodology for crawling is presented later.

Having four MWTs together as one single sentence would lead to a wrong analysis and tagging. In order to prevent this problem to happen, a period was added before any closing HTML tag⁴⁹ and before `
` or `
` so that any list item had a printed sentence boundary. Any consequent excess of periods is removed after tokenization.

- c. **Clean text up.** Any script embedded in HTML code as well as the HTML code itself was removed. Likewise, HTML entities were converted to ANSI characters so that plain text is left.
- d. **Non-relevant language deletion.** As previously described, some Spanish sites have some text in English (or in other languages). In Spanish files, any text in a language other than Spanish was manually removed.
- e. **Tokenize.** Tokenization was carried out to allow for further processing, but also to produce a cleaner output for token and type count.
- f. **Lower case.** In many cases, websites feature a rather liberal use of language. It is the case of capitalization, which not necessarily follows the rules of each language. For the tagging task, capitalization is important to annotate proper names and for named-entity recognition. However, given the high irregularity in the use of upper case in our corpus, capitalization turns into an issue instead of an advantage. Therefore, every uppercase letter was converted to lower case.
- g. **Tag with parts of speech.** For the part-of-speech (POS) tagging task, the TreeTagger (Schmid, 1994) was used.
- h. **Tag with Wordnet supersenses.** Nouns and verbs of the English corpus were annotated with 41 broad semantic categories (Wordnet supersenses) using the SuperSenseTagger (Ciaramita and Altun, 2006). A previous study (Burgos, 2009) shows that the tagger reaches an 89% of precision in automotive engineering texts. The tagger is not trained for Spanish yet, so no semantic tagging was carried out for Spanish.

⁴⁹ With exception of the tags for bold (``), italics (`</i>`) or underlined (`</u>`).

- i. **Extract Nominal MWT candidates.** Quiroz's (2008) syntactic patterns⁵⁰ (34 for English and 86 for Spanish) were used to extract nominal MWTs from the corpus. MWTs are extracted along with their POS and semantic annotations. Extraction started with the longest syntactic patterns and finished with the shortest ones. This assured that the shortest patterns that were subsumed in the longest ones did not retrieve nominal MWTs when they had already been extracted as part of a longer pattern. As an illustration, in the next examples, we want the chunker to extract the nominal MWTs at the right using the patterns at the left:

- 1) {<JJ.*><JJ.*><JJ.*><N.*><N.*>} → *anatomical lumbar telescopic expansion panel*
 2) {<JJ.*><N.*><N.*>} → *automatic shift knob*

But we do not want our chunker to extract *anatomical lumbar telescopic expansion panel* with the pattern in 1) and then *telescopic expansion panel* with the pattern in 2), when 2) is part of 1). In other words, any nominal MWT parsed and matched with a syntactic pattern is not parsed again with a shorter pattern.

- j. **Extract artifact MWTs.** Out of the total of nominal MWTs, those semantically tagged as *artifacts* with the SuperSenseTagger were extracted. It was also observed that most of the MWTs tagged as *person*, were actually artifacts. Therefore, MWTs tagged both as *artifact* and as *person* were grouped in a single set of artifact MWTs.

Facts and figures of text in BODY

Table 9, shows some relevant figures derived from the methodology outlined above. As for the corpus size, it can be observed that there is a huge difference between English and Spanish, even though the number of crawled sites is similar. However, the analysis of types, that is unique words, shows very similar figures for both languages. The token/type ratio for both languages suggests sites with very repetitive content.

⁵⁰ See Annex 1

	Eng	Spa
Websites	30	36
Tokens	125,297,255	11,082,949
Types	21,105	20,779
Nominal MWT candidates	13,286,069	457,702
Unique Nominal MWT candidates	13,939	10,582
Artifact MWTs	19,374	6,848**
Unique Artifact MWTs	4,610	6,848*

Table 9. Main figures of text in BODY section

There is also a similar trend in the figures for nominal MWTs, that is, a big difference between languages in the number of occurrences of nominal MWTs but a small difference in the number of unique nominal MWTs when both languages are compared.

A more interesting observation for the present research is the fact that, for English, out of the total of unique nominal MWT candidates, a 33.07% are artifact MWTs. For Spanish, on the other hand, no semantic tagging was possible to automatically determine artifact MWTs. However, 6,848 unique artifact MWTs were manually identified; this means a 64.71% out of the total of unique nominal MWT candidates for Spanish.

3.3.2.2. Image analysis

In order to confirm the relevance of our search space with regard to the visual component of the BC, 4 sites in English and 8 sites in Spanish were analyzed⁵¹. The methodology below shows how these sites were spidered for image information and how data was processed:

⁵¹ Technical and time constraints did not allow to obtain the image information from all the sites in our search space

1. **Extract URLs from crawler output.** A list of URLs to be spidered for image information was built from the output of the crawler used to obtain data for nominal MWT analysis (see §3.3.1).

2. **Spider Web pages for image information.** Every single URL was spidered to collect information of images.

3. **Mine image information.** Image path, name, size and file type were extracted from the output of the spider.

4. **Sort and remove duplicates.** A total of 434,065 paths to images were collected for English and 1,835,030 were collected for Spanish. After sorting and removing duplicates, 17,669 unique paths remained for English and 12,436 remained for Spanish. The high number of duplicates could be explained by the fact that most of the pages iteratively point to the same images when such images are used for the webpage layout (buttons, banners, backgrounds, logos, etc.).

5. **Analyze data.** Images were counted, their predominant type was defined, and an average size for relevant images was calculated. Relevant images were identified through a visual and manual evaluation. The average size for relevant images was 16kb with a standard deviation of 7.17. This calculation was carried out on 22,957 relevant images of two sites, one of each language (16,109 images for the English search space and 6,848 for the Spanish search space).

5. **Characterize relevant images.** As for the image size, considering an average size of 16kb and a standard deviation of 7.17, sizes between 12 Kb and 20 Kb seem to cover a representative number of relevant images. On the other hand, with regard to image name, it was observed that publishers do not name their images after a spare part name (e.g., alternator.jpg) since it would turn unmanageable in large image collections. Therefore, alphanumeric codes, sometimes including hyphens, are mostly used. Likewise, the preferred image file type for relevant images is JPG.

Facts and figures of images

	eng	Spa
Websites	4	8
Images	17,669	9812
Relevant images	13,024	6,848
Average size (all images)	14.1kb	22.5kb
Average size (relevant images)	12.7kb	24.1
All languages average size (relevant images)	16kb	

Table 10. Main figures of images in the search space. Counting is presented after removing duplicates.

3.4. Attained goals and drawbacks

Some conclusions can be drawn from the observations and analyses performed in this chapter. Two main lines can be followed to outline such conclusions: the attained goals related to the appropriateness of the search space to produce instances of the BC and the drawbacks inherent to online documents which hindered some tasks.

3.4.1. Attained goals

The text analysis here aimed at verifying the occurrence in the search space of a representative number of artifact MWTs and artifact images so the BC hypothesis model can be successfully applied. Even though not all the websites of the ODP could be crawled and analyzed, the goals were reached. The analysis of keywords and descriptions (the only one done in all the websites) provided a close and encouraging overview of the contents of the search space. It confirmed the higher variability of subdomains in the Spanish search space and the clear delimitation and specificity of the English search space, as suggested by the categories selected in the ODP for our observations.

The next step consisted on analyzing the BODY section of the HTML documents. However, it was clear that not all the webpages of relevant websites are relevant for this study. Therefore, an analysis of the strings in the URLs of a sample of websites was carried out in order to make decisions related to the optimization of web crawling. The objective of this analysis was to determine the patterns followed by the URLs which tend to present instances of the BC. It was observed that those URLs with strings like *shop*, *products*, *store* and *cataloguesearch* tend to maximize the occurrence of BC instances. Accordingly, crawling and analysis were limited to webpages with such strings in their URLs.

The BODY contents were cleaned up and processed, and nominal and artifact MWTs were extracted and counted. The results show representative number of nominal as well as artifact MWTs. It could be observed that the number of artifact MWTs reported here is lower than the number of images in the same search space mainly because we used long syntactic patterns for extraction; the shortest pattern for extraction was made up of three parts of speech. This is a decision which certainly benefits of avoiding noise, but, on the other hand, some shorter image descriptors might also been lost during extraction.

The analysis of images is also satisfactory. 19,872 relevant images were manually verified, which could mean a higher number if those images that could not be verified were to be included. Such a good sample of relevant images has been very valuable to define the prototypical image for our BC model. The downside of the observed images is the big variability in their size. An average size and a size range were determined as representative of relevant images, but it seems the sample do not present a normal distribution so the selected size range might still exclude other relevant images of smaller or bigger size. Image names, on the contrary, seem to follow a distinctive regular pattern of alphanumeric codes which could be used to identify relevant images in large image collections.

As a result of the characterization of these two components of the BC, a kind of function can be profiled in the sense that one can hypothetically assign one or more elements of the artifact MWTs set to one or more elements of the artifact image set. And that is the essence of the BC hypothesis model that will be concretized later in other chapters.

3.4.2. Search space drawbacks

As a segment of the Web, the ODP inherits its parent's properties, characteristics and limitations. The difference and advantage of choosing the ODP as our web segment, as said before, is the organization and category annotation provided by a directory structure. However, at a finer level, websites and webpages pose the same challenges of pages and sites in the whole Web. Therefore, apart from aspects related to directory structures, the analysis and strategies here could be applied to any other segment of the Web but also will find the typical problems of dealing with information in the Web.

As part of the Web, the ODP is susceptible to the problems listed by Baeza-Yates and Ribeiro-Neto (1999, p. 368) concerning the data in the Web. Broadly speaking, those problems are related to data distribution, high percentage of volatile data, large volumes of data, unstructured and redundant data, quality of data, and heterogeneous data. It is expected, however, that being the ODP a human made directory controlled by means of quality guidelines, such problems can be minimized in some degree. See below some specific examples of these problems observed in our analysis.

Capitalization and punctuation. With regard to the data analyzed in keywords and descriptions of websites, there are, for instance, misuse of capital letters and omission of punctuation. This represents an issue specially for part-of-speech tagging and for chunking since capitals are interpreted by taggers as proper names and punctuation is used for sentence or phrase boundaries.

There seems to be also a misunderstanding by some website developers of the way search engines work. This is reflected by the fact of including the same word in lower and in upper case as well as in singular and plural as well as with and without accent; that is, the same word several times with different forms. See, for example, this case of the word *transmisión* which was found with four different forms in the description of a Spanish website: *Transmisión transmisión Transmision transmision*. Likewise, there are no punctuation marks between them.

Redundancy. As shown in the analysis of the BODY section (§3.3.2.1.2), text in the search space is highly repetitive. It can be understood and could be predicted due to spread layout for online

product catalogs. This rather general layout tends to have static sections with company information and product categories and a dynamic section with product information. Therefore, no matter what product the user is visualizing, the static content is going to be displayed again and again for every single page.

Inaccessible data. With regard to the format, it also has to be said that the number of Web pages containing PDF documents or flash movies embedded in the HTML code is considerably increasing. This poses an additional problem since access to contents in these formats is not straightforward. On the other hand, besides the aesthetic aspect, sometimes the use of these formats also seems to show the intention by the site's publisher not to be indexed or crawled beyond the simple meta tag values for keywords and descriptions.

Codification. Heterogeneity of pages codes (Unicode, ANSI, etc.) in websites caused some characters to be downloaded as non-text characters (squares of triangles). This made some of our scripts or applications crash when processing files.

Other languages. Particularly for the Spanish search space, some sites have contents in other languages, especially English, even when it is not a bilingual or multilingual site. This also caused problems to be fixed before tokenization, tagging and chunking.

Non-standard syntax. Sometimes because of space constraints and other times due to careless use of language, some MWTs in both languages do not follow the standard syntax rules. Let us examine some few examples:

English

- a. *Air/Fuel Ratio Calibrator ARC2-A* → ARC2-A Air/Fuel Ratio Calibrator*
- b. *Air Filter Element 3.0CSi-3.0CSL* → 3.0CSi-3.0CSL Air Filter Element*
- c. *Boot Lid Rubber Buffer BMW* → Boot Lid Rubber Buffer for BMW*

Spanish

- d. *Relé intermitencia 12v c.a. 3 terminales → Relé de intermitencia de 12v c.a. de 3 terminales*
- e. *Reten 22+32+5.5 arranque Lambretta* → Retenedor 22+32+5.5 de arranque para Lambretta*

The recurrent irregularity for English seems to consist of one or several MWT components misplaced as head nouns. The right place for such *pseudo head nouns* should be as premodifiers, as is the case of examples *a* and *b*, or as postmodifiers, as in example *c*.

For Spanish, on the other hand, the dominant irregularity seems to be the omission of prepositions, as in examples *d* and *e*, or even some morphological changes as the shortening of *Retenedor* in example *e*.

The imperfections above certainly affect the quality of our search space and demand additional processing tasks to fix them or at least to handle them. Their identification in this characterization process, however, contributes to a more effective performance of further tasks at later stages of this thesis.

4. NOMINAL, ARTIFACT MWT RECOGNITION

4.1. Introduction

In previous chapters, we introduced and described the bimodal co-occurrence model (see Chapter 1) and characterized the search space where instances of the BC model typically occur (Chapter 3). It was shown how an image and a term can be independent representations of the same real world entity and that it is the meaning coincidence of these two different modes that constitutes the BC hypothesis. We suggested then that matching two images of the same artifact in two different documents would make possible their respective denominations match too and that a recursive exploitation of the model would enable monolingual, bilingual or multilingual dictionary generation.

This chapter proposes a methodology towards the image-term (or denomination) alignment task. The procedure for this task takes into account possible variations in document layouts. That is, there are some ideal web layouts where the unique text surrounding the image within reasonable boundaries is the artifact's denomination. In such a case, a rather simple algorithm aligns the term with its image. However, sometimes there are considerable amounts of text surrounding the image. In this case, the text must be parsed and the artifact's denomination must be recognized among all the term candidates in the text. For this latter scenario, nominal multi-word term (MWT) recognition is performed to extract relevant units and narrow down the list of candidates (see §4.2 below).

However, even when nominal MWT recognition certainly helps reducing the search space into a candidate list, the artifact's denomination still needs to be selected. For this, two approaches are followed, namely, anchor-based selection and noun classification. The former uses alphanumeric codes as anchors that consistently appear next to artifact MWTs in catalogs. The latter classifies head nouns in a nominal MWT list into concrete and abstract nouns with the assumption that concrete nouns maximize the probability of being the term for a given image (see §4.3 below). After noun classification, the image is annotated with a short list of artifact denomination candidates.

Below, the approaches explored here to tackle the above mentioned problems are described. Then, the evaluation and global figures of each task are presented.

4.2. Nominal MWT recognition

Before going on to describe and evaluate the approaches proposed here for nominal MWT recognition, let us remind the way we defined nominal MWTs in Chapter 1 (§1.2.1.2): Nominal MWTs are lexical items that: (a) can be decomposed into multiple lexemes; (b) display semantic idiosyncrasy; (c) have a noun as their nucleus; and (d) are terminological units used in special subject fields (Cabré, 2000).

While this definition accounts for the theoretical delimitation of nominal MWTs, the practical recognition of these lexical items implies additional necessary considerations which not only affect the outcomes of the proposed approaches but also the evaluation of their performance. Such considerations are discussed below.

4.2.1. Preliminary considerations

In order to predict the approximate performance of machines in the delimitation of MWTs, it could be said that: a) if humans, using their cognitive and linguistic competences, were capable of accurately delimiting MWTs, and b) if such competences could be hypothetically transferred to a computer, then it could be assumed that computers should perform in the task as precisely as humans.

However, the reality is that even humans do not agree on what are or what should be the boundaries of a term. Therefore, the same degree of uncertainty is expected from machines. In a study by Estopà et al. (2007), it was proven that even experts present variance when delimiting the terms of their own specialty area. It was observed that there is certainly more agreement to detect windows where the nucleus of the term was present, but the precise delimitation of the terms was not that unanimous:

[...] we are far away from the desired unambiguous term identification and delimitation; even specialists, who have been traditionally considered competent for the task, have problems with it. [our translation] (Estopà et al., 2007)

When detecting specialized knowledge, it is the cognitive competence that enables an expert to do so. When delimiting a term, on the other hand, the linguistic competence plays an

important role. That is the reason why human translators, who are expected to be more linguistically aware, perform as well as experts – and sometimes even better – when delimiting terms. But even translators or linguists are not exempt from disagreement when facing this task.

There could be a great deal of reasons causing difficulties in term delimitation. Given the relevance for our experiments, however, we will focus on some features that fall on the grounds of linguistic knowledge (e.g., morphological, lexical and syntactic). Such knowledge will be used to determine the formal configuration that a term candidate should have so it increases its chances to be a MWT.

4.2.1.1 Definition of formal features for MWT recognition

The formal features for nominal MWT recognition consist of a) the part-of-speech categories that are allowed to be part of MWTs, and b) the items that are permitted in each allowed category. Term candidates with length equal to or greater than two words will be extracted according to these features.

4.2.1.2 Allowed part-of-speech categories

For English, as described and justified in Chapter 1 (§1.2.1.2.1), only MWTs with premodification are included in this study. Accordingly, the predefined part-of-speech categories for English nominal MWTs are (Penn-Tree-Bank tag set used⁵²):

- Common nouns (NN, NNS)
- Adjectives and past participles in adjectival function (JJ)
- Proper nouns (NP, NPS)
- Adverbs ending in *-ly* (RB) when modifying an adjective that modifies a noun.

For Spanish, the predefined part-of-speech categories for nominal MWTs are (TreeTagger tag set used⁵³):

- Common nouns (NC)
- Adjectives in post-modification position (ADJ)
- Past participles in adjectival function (VLadj)

⁵² <http://www.cis.upenn.edu/~treebank/>

⁵³ <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

- *de* preposition (PREP)
- Adverbs ending in *-mente* (ADV) when modifying an adjective that modifies a noun
- Proper nouns (NP)

Some of the decisions on the configuration of these lists of predefined categories for both languages were made on the base of the characterization of the search space carried out and described in Chapter 3. According to this characterization, there are a number of technical and language use features that make our search space rather noisy. Among these features, the most relevant ones for the term recognition task are the frequent non-standard use of language syntax, abbreviations and capitalization, and the extensive use of numbers due to references to product codes, prices, measurements, etc. This is the reason why proper nouns (NP) are included in the predefined categories and numbers are excluded. For example, a lot of terms including common nouns are frequently capitalized, which makes the tagger categorize them as proper nouns. Likewise, even when numbers may sometimes be part of MWTs, the high frequency of non-terminological numbers in our corpus seems to make the exclusion of numbers from predefined categories more beneficial than harmful.

It has also been observed that adverbs are hardly found to form part of MWTs with the exception of derived adverbs of manner ending in *-ly* for English and in *-mente* for Spanish. It is also important to restrict these allowed adverbs to those that appear modifying an adjective and not modifying a verb when the term is the subject of the sentence:

- ◆ *the Continuously Variable Transmission (CVT) proved 35% more efficient than the Manual Transmission (MT)*
- ◆ **image registration usually consists of four major steps...*

In the examples above, the first instance underlines an allowed adverb and the second one underlines an excluded adverb.

As for the position of adjectives in Spanish, empirical observations suggest that, most of the times, adjectives in MWTs appear in post-modifying position. This trend has an explanation from the grammar and semantics of Spanish adjectives. As described by the Real Academia Española (252ff.), when the adjective is in post-nominal position it is said to be restrictive, relational, and classifies the noun. When it appears in prenominal position, it mostly implies

adverbial, determinant, affective, or evaluative value. Likewise, preposed adjectives increase ambiguity and figurative senses. Therefore, only postposed adjectives are allowed in our categories here. For example, in cases like *nuevo marco regulatorio*, the adjective *nuevo* will not be considered part of the MWT. The value attributed by its prenominal position makes it irrelevant to this term. It could be certainly moved by some speakers to a post-nominal position (*marco regulatorio nuevo*) which would be still grammatical, but it could also affect the semantics of the noun phrase. The problem is not trivial and has triggered interesting discussions by a number of grammarians, as reviewed by Whitley (2002, p. 230-236). We are not going deeper in this discussion, though, but will try to take advantage of the fact that it is much less probable to have a specialized adjective in prenominal position (*nuevo regulatorio marco**, *regulatorio marco nuevo**) than in post-nominal position.

With regard to prepositions, an analysis of 62,767 entries in three Spanish specialized dictionaries shows that *de* appears in 46.5% of the MWTs. The other prepositions all together account for 12.8% of the entries (see Figure 23).

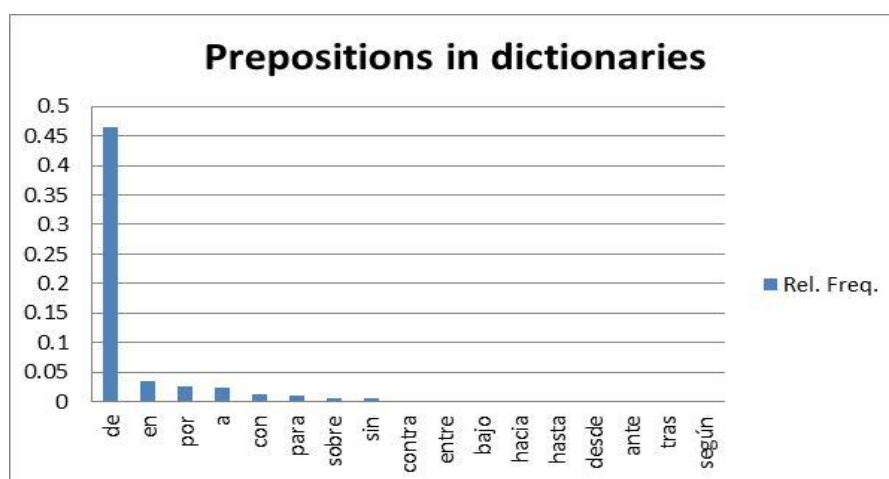


Figure 23. Analysis of prepositions in dictionaries

Therefore, *de* is the only allowed preposition for our formal configuration or our MWTs in Spanish not only because of its representativeness in dictionaries, but also because of its inherent capacity to modify other nouns when leading off a prepositional phrase (*planes de adquisición gradual de derechos*). Although *de* can play other roles, its high modification power makes it the most recurrent preposition used in Spanish MWTs.

Last, in relation to articles, the grammatical function of articles, their semantic implications,

their low representativeness in dictionaries (7,032 – 11.2%), and the noise generated when including them in patterns during experimentation led us to exclude them from the list of our Spanish predefined categories of MWTs.

Taking into account the frame and the restrictions of this formal definition as well as the related discussion in Chapter 2 (§231), three approaches for nominal MWT recognition are explored here, namely, a rule-based method (also used here as baseline), a seed-based bootstrap method, and a predefined-categories-based method. For these three recognition methods, we depart from the fact that nominal MWTs are embedded in noun phrases. Likewise, the methods followed here are based on the assumption that given a specialized text, noun phrases increase their probability of containing or even being MWTs at the same time. Accordingly, we first focus on noun phrase recognition and extraction which seems to be a lower level task and then check termhood by measuring precision and recall.

In order to be able to work with part-of-speech categories and to alternate lemmas and forms, the observation and test corpora were tagged and lemmatized using the TreeTagger (Schmid, 1994). Below each method for MWT recognition and their results are described, discussed, and evaluated.

4.2.2. Rule-based MWT recognition

It can be generalized that the syntax of a nominal MWT is that of a noun phrase. This fact has enabled a good deal of research on term recognition using symbolic or linguistic methods (Savary & Jacquemin, 2003; Vivaldi & Rodríguez, 2007; Estopà, 1999; among others). However, references in the literature to nominal MWTs have been mainly limited to short noun compounds. Even for noun phrases, studies had been limited to very short constructions, which motivated Quiroz's dissertation on English-Spanish complex noun phrases (Quiroz, 2008). Similarly, considering that nominal MWTs follow the grammar and the syntax of their language system, it is expected that complex nominal MWTs also occur. See for example the MWTs below in both languages taken from a professional translation:

eng: *auto anti-glare inside rear view mirror*

spa: *espejo retrovisor interior antideslumbramiento automático*

eng: *sequential multiport fuel injection system*

spa: *sistema de inyección secuencial de combustible de orificios múltiples*

Quiroz's syntactic patterns account for the complexity that our MWTs could feature. Therefore, such patterns are used here as a departing point for our rule-based term recognition, although they are tuned up for a better performance during term recognition. Adjustments to Quiroz's patterns are carried out to create rules of a shorter length (Quiroz's are equal or greater than 3 words) and to include and/or exclude categories and items according to the formal configuration of MWTs defined above. The final version of the syntactic patterns used here and defined as baseline for our experiment can be seen in the Annex 1.

After adjusting Quiroz's syntactic patterns, they are put in a chunker (22 for English and 70 for Spanish) and used to recognize and extract nominal MWTs from text annotated with part-of-speech tags. Any tag sequence matching our patterns is considered a term candidate and therefore extracted. The patterns in the chunker were sorted by their length starting from the longest one, so the extraction starts with the longest syntactic patterns and finishes with the shortest ones. This strategy assures that the shortest patterns that are subsumed into the longest ones do not retrieve nominal MWTs when they have already been extracted as part of a longer pattern. As an illustration, let us use a previous example again. We want our chunker to extract the nominal MWTs at the right using the patterns at the left:

JJ + JJ + JJ + NN + NN → *anatomical lumbar telescopic expansion panel*
JJ + NN + NN → *automatic shift knob*

But we do not want our chunker to extract *anatomical lumbar telescopic expansion panel* with the first pattern and then *telescopic expansion panel* with the second pattern, when the second is part of the first. In other words, any nominal MWT already parsed and matched with a syntactic pattern is not parsed again.

4.2.3. Seed-based bootstrap MWT recognition

As suggested above, the rule-based approach restricts recognition and extraction to a set of predefined rules or syntactic patterns. Any set of rules, however, risks missing relevant patterns that were not foreseen when defining the rules, among other reasons because there are not theoretical limits for English or Spanish piling up modifiers.

In order to relax such restrictions and to allow for non-foreseen patterns to be accounted for, a seed-based approach and some variations of it have been proposed in the literature, as was shown in Chapter 2 (§2.3.1). As an alternative to rule-based MWT recognition, a seed-based

bootstrap method is also adopted here (following Nazar et al., 2012, and Baroni & Bernardini, 2004) with adaptations motivated by the observations and analyses made so far on our terms, dictionaries and search space.

Our seed-based approach operates on tagged text and alternates with tags, lemmas and forms in an attempt to reduce the error generated by the part-of-speech tagger. Even when lemmatization necessarily implies part-of-speech tagging, working sometimes on lemmas and forms rather than on tags may help improve performance in recognition. For example, the MWT *bar pivot bushing fix cap*, where *bushing* is a noun, was tagged as: *bar/NN pivot/NN bush/V fix/NN cap/NN*.

If we rely only on tags, two terms *bar pivot* and *fix cap* would have been identified here because *bushing* was tagged as verb which is considered a term boundary here. If we work on lemmas, however, given that *bush* is also a seed in our noun list, the term would not have been split, the boundaries would have been looked for beyond *bar* and *cap*, and *bush* would have been turned back into its original form *bushing* so the MWT is conserved in its original form.

For Spanish, a special instruction was also defined with regard to past participles in adjectival position (VLadj). For this category, the form is used instead of the lemma because the tagger lemmatizes it as a verb and not as an adjective, which would cause it to be stopped in case it matches a verb included in the verb list.

Likewise, the *<unknown>* cases had to be handled. The TreeTagger assigns the lemma *<unknown>* when the word is not included in its lexicon. This does not avoid, though, that the word is annotated with a part-of-speech tag. Therefore, if the assigned tag falls in the list of predefined categories, the form is kept.

Seed recognition and expansion

The whole procedure for seed recognition and expansion consists of two phases: a) use of nouns as seeds to extract a first list of terms, and b) bootstrap term extraction using modifiers of terms in the phase-1 list.

4.2.3.1. Noun-based recognition and extraction

For this phase, we built a seed list of 29,891 English nouns and 20,409 Spanish nouns manually extracted from three different sources types: *Diccionario de la Automoción* (South & Dwiggin, 1999), EuroWordNet (Vossen, 1998), and three online automotive specialized catalogs. After sorting and removing duplicates, a final seed list of 11,177 English nouns and 12,156 Spanish nouns was obtained.

Given that we used nouns from EuroWordNet, which also contains general nouns, these lists are a mix of general and specialized nouns. Instead of hurting, however, this mix is beneficial since around 85% of MWTs head nouns are in the intersection of specialized dictionaries and general dictionaries, as shown by an analysis in the Introduction chapter (§1.2.1.2). We expect that including nouns from additional specialized sources to the lists improves the assumed 85% coverage of general language nouns. It should also be emphasized that no MWTs were included as seeds in the lists. It is expected that expanded seeds help recognizing MWTs of various lengths and syntactic configuration that may not be included in dictionaries and that may have not been predicted with the rule-based approach.

The algorithm for this phase can be described as follows: given a list of seed nouns, read each line of the corpus looking for each of the seeds. For each found seed, span a window leftward and rightward until a boundary is found. Then, extract the expanded seed.

The boundaries to stop seed expansion are determined by the first occurrence at right or at left of any category not included in the predefined categories defined above in §4.2.1. Such non-allowed categories also include:

- A new line
- Any non-alphabetic character including punctuation signs (with the exception of hyphens). For this study, even commas are considered delimiters. Let us consider this example: *fast-burning, efficient combustion system*. A comma may act as a coordinating conjunction suggesting that two terms share the same head noun (*fast-burning combustion system* and *efficient combustion system*), but it also can separate inherent modifiers (rightmost) from external modifiers (leftmost) of the same term. In the former scenario, the right decision would be made when discarding the

leftmost modifier. In the latter case, we would be certainly discarding one term candidate, but resolving coordination is a problem out of the scope of our current task of term recognition.

- English unambiguous verbs. We include in this stop set 4,578 verbs that rarely work as nouns or as adjectives. This unambiguous verb list was obtained by using a noun and an adjective list to filter a general list of English verbs.
- Evaluative adjectives. This category consists of a small set of subjective adjectives such as *good*, *excellent*, *interesting*, etc.

If it happens that the boundary is right after the seed but not before it, then the seed is actually the nucleus of the term. Likewise, if it happens that the boundary is right before the seed but not after it, then the seed could be a modifier of the head noun or of other modifiers. And if it happens that there are boundaries right before and after the seed, then it would be a one word term candidate. For instance, the seed *bracket* would be expanded leftwards if the candidate is *alternator mounting bracket*, or rightwards if the candidate is *bracket lock nut*, or even in both directions if the candidate is *alternator mounting bracket lock nut*.

4.2.3.2. Modifier-based recognition and extraction

This second phase is a bootstrapping operation which dynamically uses the modifiers of the terms generated in the phase 1 as seeds. The algorithm for this phase can be described as follows: given the term list generated in phase 1, build a new seed list with any word not included either in the original seed list or the stop list used in phase 1. Given this new seed list, read each line of the corpus looking for each of the seeds. For each found seed, span a window leftward and rightward until a boundary is found. Then, extract the expanded seed. For example, the seed *impulse* would be expanded leftwards if the candidate is *point target impulse*, or rightwards if the candidate is *impulse response measure*, or even in both directions if the candidate is *point target impulse response measure*.

4.2.4. Predefined-categories-based MWT recognition

This approach relaxes restrictions even more since it only bases on tags to extract every sequence of words whose tags fall in the set of predefined categories (following and adapting Bourigault, 1992). The mere use of predefined categories seems *a priori* to generalize more and

to cover the previous method but it is also expected to generate more noise.

The algorithm for this approach can be described as follows: given a list of predefined categories, extract every word whose part-of-speech tag is included in the list of predefined categories and extract them in the order they appear in the corpus provided that the word is not included in the stop list and that the resulting sequence has the required length (≥ 2). See for example the lemmatized sentence below:

this/DT report/NN describe/VBZ an/DT initial/JJ assessment/NN of/IN the/DT absolute/JJ geolocation/NN accuracy/NN of/IN the/DT NASDA/NP JERS-1/NP Amazon/NP mosaic/NN image/VBN during/IN low-water/NN season/NN ,/, September-December/NN 1995/CD ./SENT

The application of the algorithm yields these candidates: *initial assessment; absolute geolocation accuracy; NASDA JERS-1 Amazon mosaic; low-water season*. The other strings in this sentence were not recognized as candidates because their tags are not in the list of predefined categories (e.g., *DT*, *VBZ*), or they are in the stop list, or they do not have the required length (e.g., *report*).

4.2.5. Evaluation of MWT recognition

This subsection provides an evaluation of the three approaches for MWT recognition just presented above (including the baseline). For this evaluation, a test set for each language was prepared. The test set consists of a number of randomly selected MWTs and a context of use for each term. The evaluation aims at measuring the performance of each method to recognize in each context what we call here *the referent MWT*. Precision and recall are also measured.

The test set for Spanish, with 403 MWTs, was taken from *Le grand dictionnaire terminologique* (GDT) that is available through the Office Québécoise de la Langue Française⁵⁴. The terms belong to the Telecommunications and Assurance domains. The test set for English, with 288 MWTs, was taken from *Termium Plus*[®], the Government of Canada's terminology and linguistic data bank⁵⁵. The terms belong to the Telecommunications domain.

⁵⁴ <http://gdt.oqlf.gouv.qc.ca/>

⁵⁵ <http://www.btb.termiumplus.gc.ca/>

The methodology for the evaluation can be outlined as follows:

- a) Build a test corpus with the contexts, each context in a new line.
- b) Lemmatize and annotate the corpus with parts of speech.
- c) Apply each recognition method to each context and extract term candidates. Methods are abbreviated as follows (lang = eng, spa):
 - a. Rule-based (baseline) = lang_rules
 - b. Seed-based = lang_seeds
 - c. Predefined categories = lang_allowed
- d) Align each referent MWT with the extracted candidates.
- e) Measure the relative edit distance between the referent MWT and each candidate. The edit distance measures the similarity between two strings. It counts all the necessary transformations needed to convert a string into another string. If no transformation is necessary, the distance is 0; if the number of transformations is equal of greater than the length of the string, the distance is 1. A relative value is given between 0 and 1.
- f) Carry out a manual evaluation of the recognition:
 - a. If the edit distance is 0 for any of the candidates, then mark it as recognized and assign a 0 as evaluation score.
 - b. If the edit distance is not 0, then evaluate:
 - i. If the referent MWT is embedded in the most similar candidate, i.e., the candidate is longer, mark it as recognized, but assign an evaluation score using a positive digit indicating the number of extra words in the candidate.
 - ii. If the referent MWT is longer than the most similar candidate, mark it as recognized, but assign an evaluation score using a negative digit indicating the number of missing words in the candidate.
 - iii. If the referent MWT was not recognized at all, then mark it as missed.
- g) Count the number of recognized and missed MWTs.
- h) Compute mean and standard deviation (σ) for the edit distance.
- i) Compute mean and standard deviation (σ) for the evaluation score.
- j) Compute analysis of variance (ANOVA) for the edit distance and for the evaluation score.
- k) Compute precision and recall.

4.2.6. Results

Figure 24 shows a comparison of recognized vs. missed MWTs by the three approaches for both languages. This figure includes counting of all the cases where the referent MWT was marked as recognized, even if the length of the most similar candidate is not the same of the referent MWT, what can be seen as a relaxed measuring of recall. That is, recognized in this first part of the evaluation means that a) the referent MWT is embedded in the candidate; or b) the candidate is a truncated form of the referent MWT; or c) the candidate is identical to the referent MWT. The ranking of the three methods according to this relaxed recall evaluation is as follows: for English, predefined categories (93.75%), seeds (87.15%), and the rule baseline (85.41%); for Spanish, predefined categories (90.81%), seeds (89.82%), and the rule baseline (75.18%). It will be shown later (§4.2.6.2) that this ranking changes when measuring standard precision and recall as well as the *F* score.

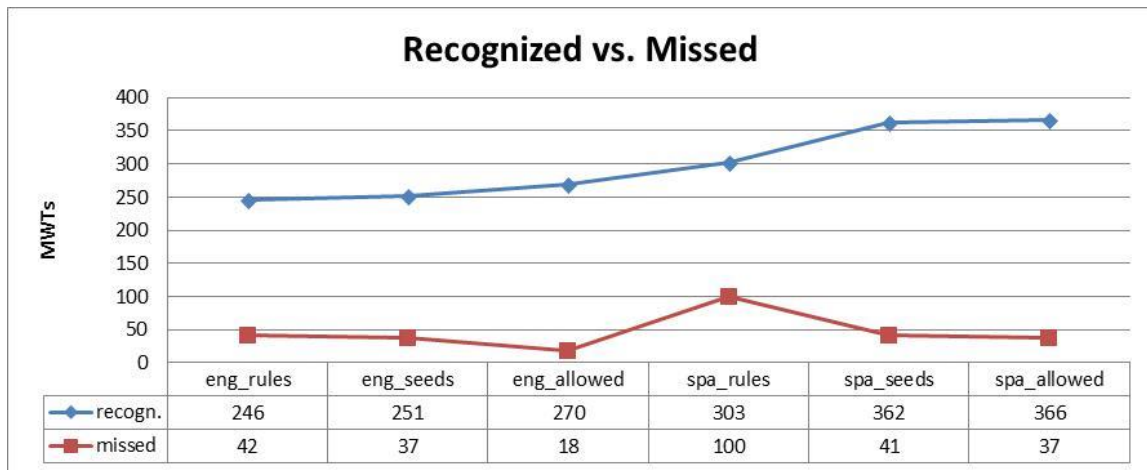


Figure 24. Recognized (blue) vs. missed (red) referent MWTs for the three methods in English (left side) and Spanish (right side).

Given that Figure 24 includes all the candidates matching the referent MWT, even if they are longer or shorter, the relative edit distance was measured and averaged in order to have a more accurate sight of the performance per method (see Figure 25). The trends as for the means are almost the same as in Figure 24, except for English where the rule-based baseline behaves better than the seed-based. The standard deviation, however, shows much less variability for the rule-based baseline.

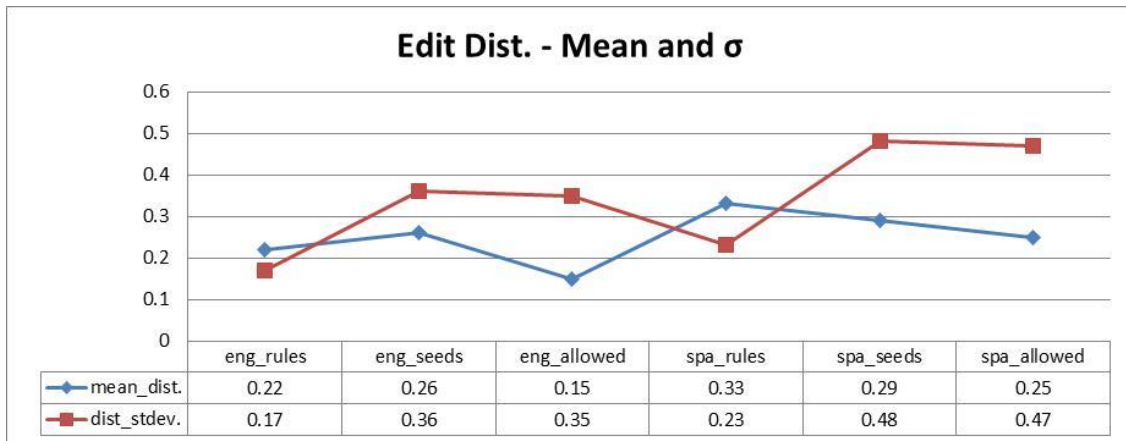


Figure 25. Edit distance mean (red) and Standard deviation σ (blue) for the three methods in English (left side) and Spanish (right side).

Another, and perhaps more accurate, perspective of the same analysis of distance between the referent MWT and the most similar candidates is shown in Figure 26. It illustrates the means and standard deviations of the evaluation score. This standard deviation confirms less variability in the length of the terms recognized by the rule-based baseline, and also starts reinforcing the trends in higher variability for Spanish.

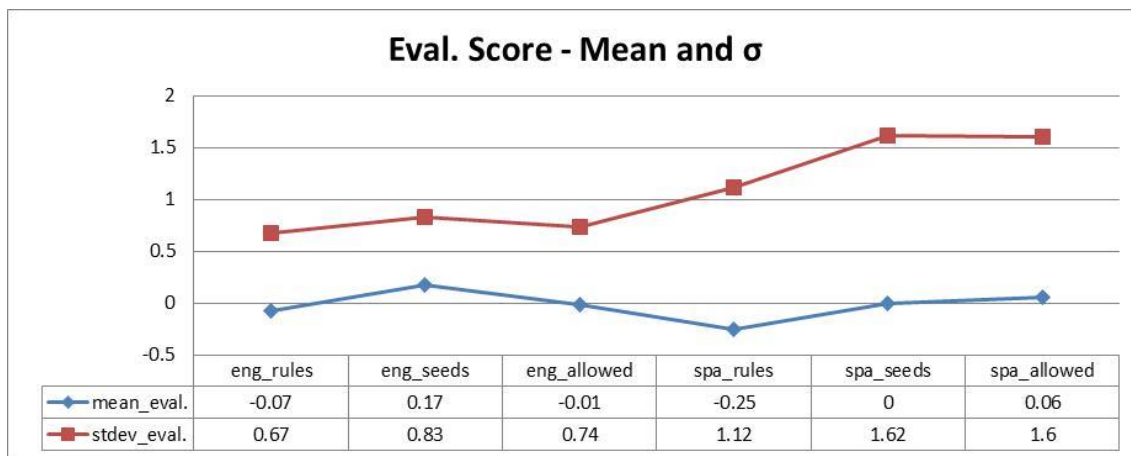


Figure 26. Evaluation score mean (red) and Standard deviation σ (blue) for the three methods in English (left side) and Spanish (right side).

To observe the previous trends in means and standard deviation from another perspective, Figure 27 shows the same measures for the number of candidates retrieved per context by each approach. It can be seen that the rule-based baseline has been showing less variability due, in part, to the fact that it is extracting fewer candidates for both languages. The higher values here for English when compared to Spanish, however, should not be taken into account since it

seems that the contexts for the English terms are longer than the context for the Spanish terms. Therefore, it is expected that the mean is higher for English and that it does not directly relate to the performance of the approaches in a single language.

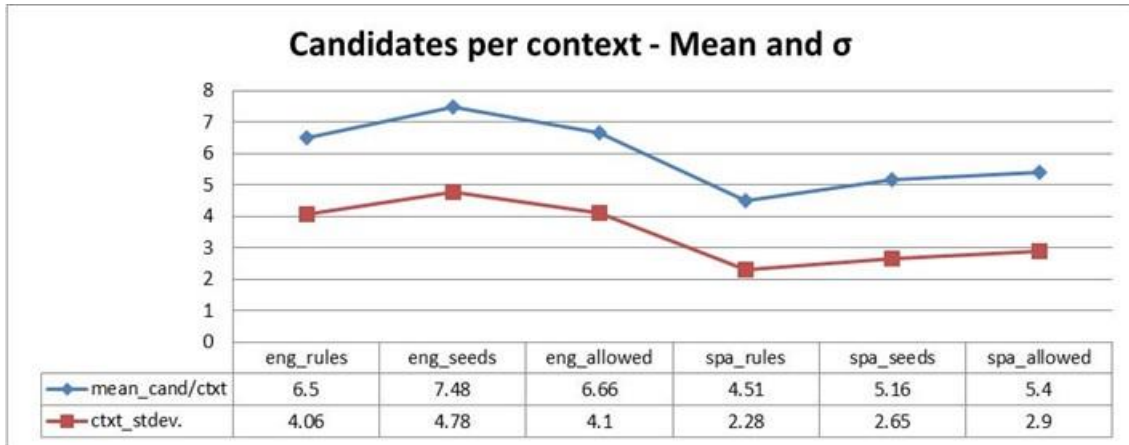


Figure 27. Candidates per context. Mean (red) and Standard deviation σ (blue) for the three methods in English (left side) and Spanish (right side).

4.2.6.1. Analysis of variance

Besides the information provided by the previous graphics and numbers, an additional analysis to determine significant statistical differences between the approaches was carried out. For this purpose, an analysis of variance was performed for each language. This time, only the evaluation score and the number of candidates extracted were taken into account for the analysis. The edit distance was left out since it proved to be less reliable for evaluation than the manual evaluation score due to the edit distance's sensibility to small changes, which are given more weight when the strings are short.

Tables 11 and 12 show the summary of the analysis for Spanish, for both variables and the three methods. It can be seen that there is statistical difference between the rule-based baseline and the predefined-categories approach as well as between the rule-based baseline and the seed-based approach. On the other hand, seeds and predefined categories perform similarly with no statistical difference. Figures 28 and 29 illustrate the differences in the medians and the variance for each method.

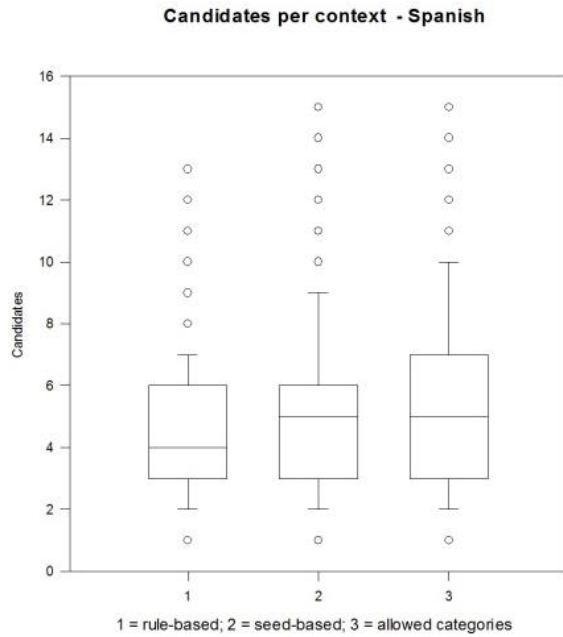


Figure 28. Candidates per context - Spanish

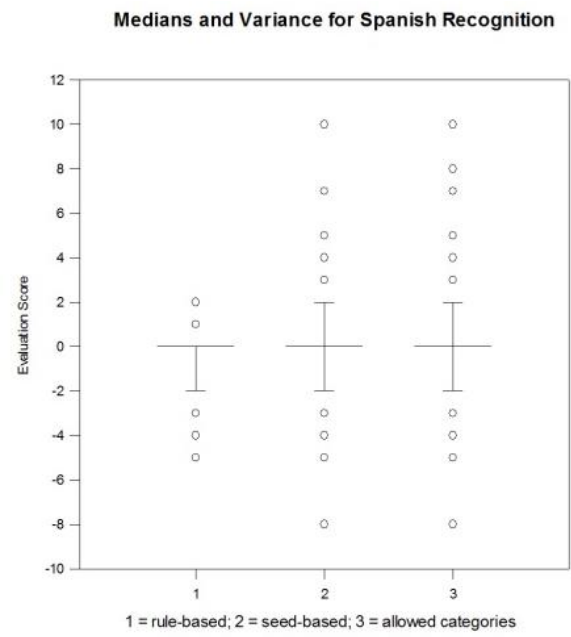


Figure 29. Medians and Variance for Spanish Recognition

Comparison	Diff of Ranks	Q	P<0.05
spa_allowed vs. spa_rules	73.158	3.163	Yes
spa_allowed vs. spa_seeds	11.153	0.506	No
spa_seeds vs. spa_rules	62.005	2.675	Yes

Table 11. ANOVA of Evaluation score - Spanish

Comparison	Diff of Ranks	Q	P<0.05
spa_allowed vs. spa_rules	84.759	3.665	Yes
spa_allowed vs. spa_seeds	18.485	0.838	No
spa_seeds vs. spa_rules	66.274	2.859	Yes

Table 12. ANOVA of Extracted candidates - Spanish

For English, however, the three approaches perform very similarly. There is only a statistical difference in the evaluation score between seeds and rules. The other combinations and variables present no significant difference as shown by Table 13 and Figures 30 and 31.

Comparison	Diff of Ranks	Q	P<0.05
eng_seeds vs. eng_rules	53.033	2.670	Yes
eng_seeds vs. eng_allowed	41.301	2.130	No
eng_allowed vs. eng_rules	11.732	0.603	No

Table 13. ANOVA of Evaluation score - English

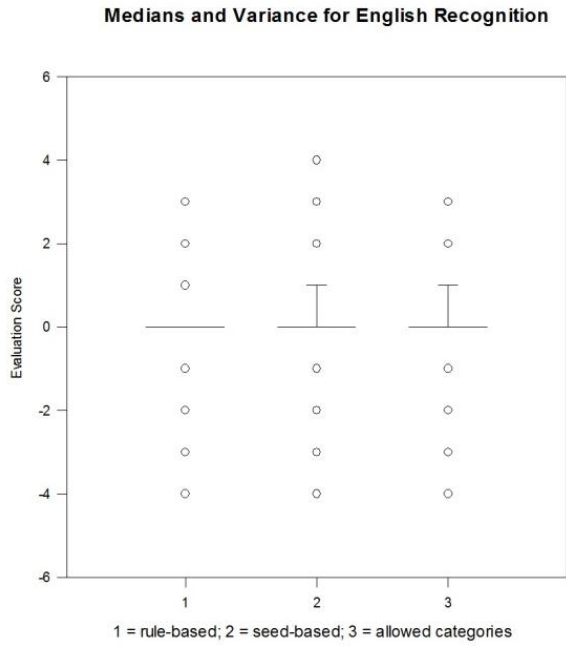


Figure 30. Medians and Variance for English Recognition

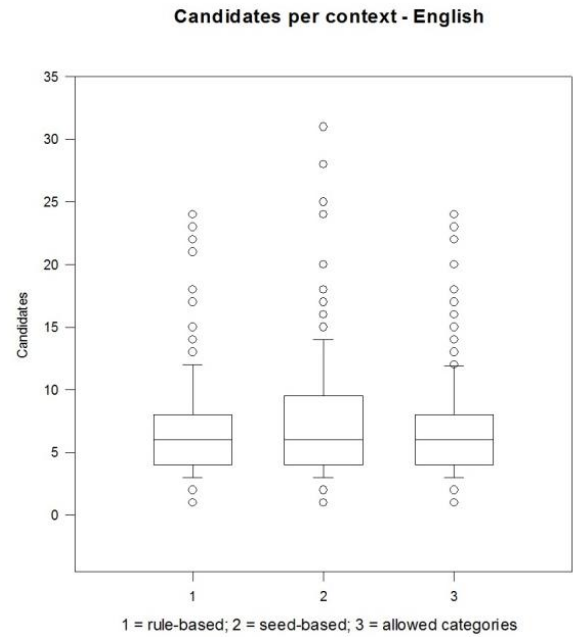


Figure 31. Candidates per Context - English

4.2.6.2. Precision and recall

Figures 32 and 33 show the evaluation of precision and recall for the three methods in both languages.

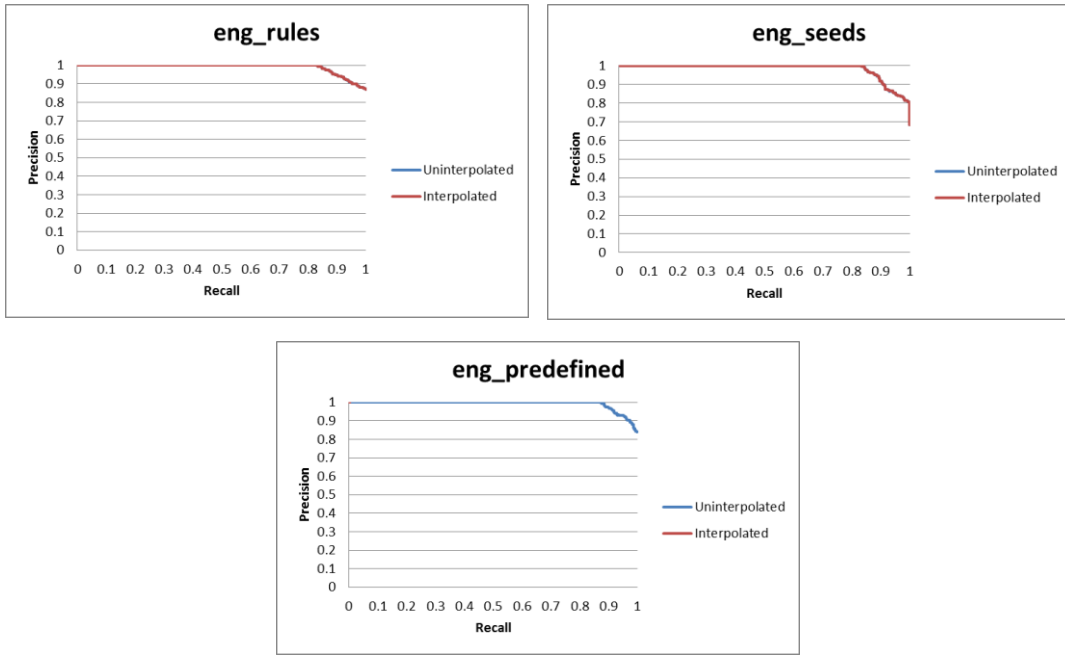


Figure 32. Precision and recall for English MWT recognition.

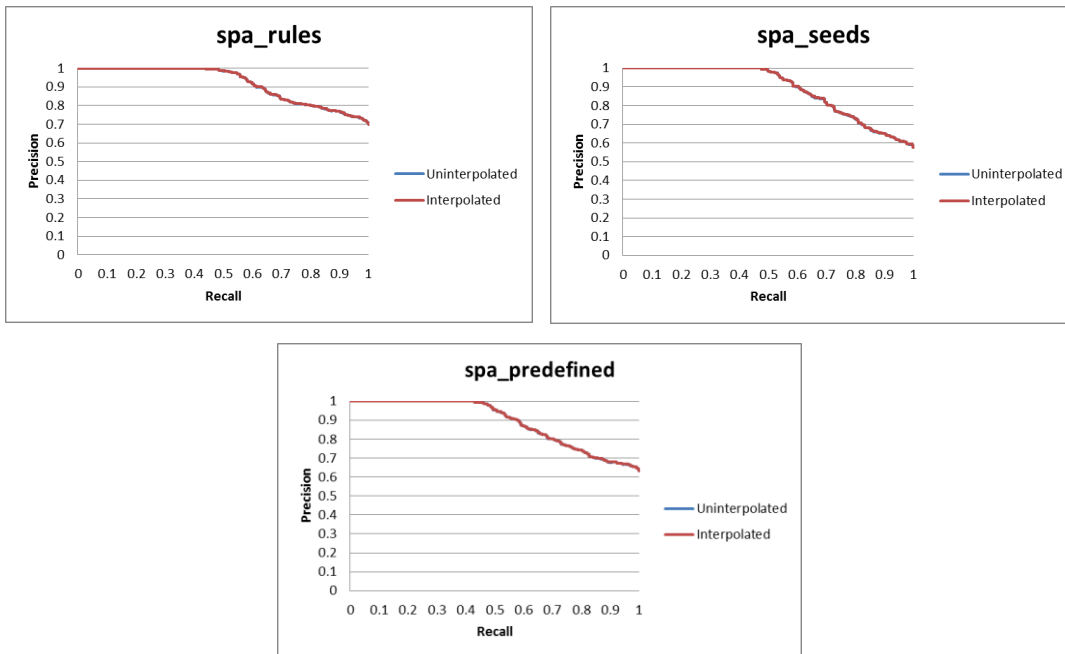


Figure 33. Precision and recall for Spanish MWT recognition

Tables 14 and 15 show aggregated scores for precision and recall reached by the three methods in both languages.

English:	Precision	Recall	F1
Rule-based (baseline)	0.87	0.85	0.86
Seed-based	0.68	0.87	0.77
Predefined categories	0.84	0.94	0.88

Table 14. Precision and recall for English MWT recognition.

Spanish:	Precision	Recall	F1
Rule-based (baseline)	0.7	0.75	0.72
Seed-based	0.57	0.9	0.7
Predefined categories	0.63	0.91	0.75

Table 15. Precision and recall for Spanish MWT recognition.

4.2.7. Discussion

The analyses in the previous section suggest some new findings and confirm some previous assumptions. The results consistently show a better performance of the three approaches for MWT recognition in the English language. Even when the seed approach seems to behave slightly better in Spanish, the variability shown by standard deviations is consistently lower in English, which compensates the little difference between languages for the seed approach. There seems to be, on the other hand, a consistency between the two languages in the variation of candidates per term (see Figure 34), which means that recall and precision can be dependent on the method but not on the language.

The fact that premodification prevails in English and postmodification predominates in Spanish for MWT formation plays an important role in the performance of these approaches. For English, no decisions have to be made with regard to prepositions and articles. These two categories are not allowed for most of the English MWTs so instead of contributing with variability, they serve as boundaries and help retrieving relevant units. In Spanish, on the contrary, prepositions and articles are part of dictionary entries. Most of the MWTs not recognized in Spanish contains non-predefined categories such as articles or excluded prepositions (*constitución de la renta, composición del activo, adquisición en bloque*). Therefore, any decision made as for the inclusion or exclusion of categories has had its consequences. Excluding articles and prepositions may truncate some terms, and including them may overgenerate pseudo modifiers.

The strategy adopted here on the predefined and non-predefined categories to be part of

MWTs, particularly in Spanish, seems to go in the right direction, though. The fact that we selected at random the MWTs for this evaluation implied that some terms were not expected to follow the patterns we defined for the formal configuration of our MWTs. This is the reason why some terms were missed, because of such decisions, although, as discussed before, the inclusion of certain categories in dictionaries is sometimes arguable from the linguistic, semantic and statistical point of view.

Let us now discuss the performance of each individual approach. The analyses consistently show differences between the rule-based baseline and the other two methods; the differences are even statistically significant, especially for Spanish. As expected at a certain extent, the syntactic patterns we defined here as baseline did not represent many of the referent MWTs in the test set. The rule-based baseline behaves well for English (F1=0.86), and almost acceptably for Spanish (F1=0.72). It is fair, however, to look at these results in the context of the total of extracted candidates (see Figure 34) and the observed sustained trends in variability.

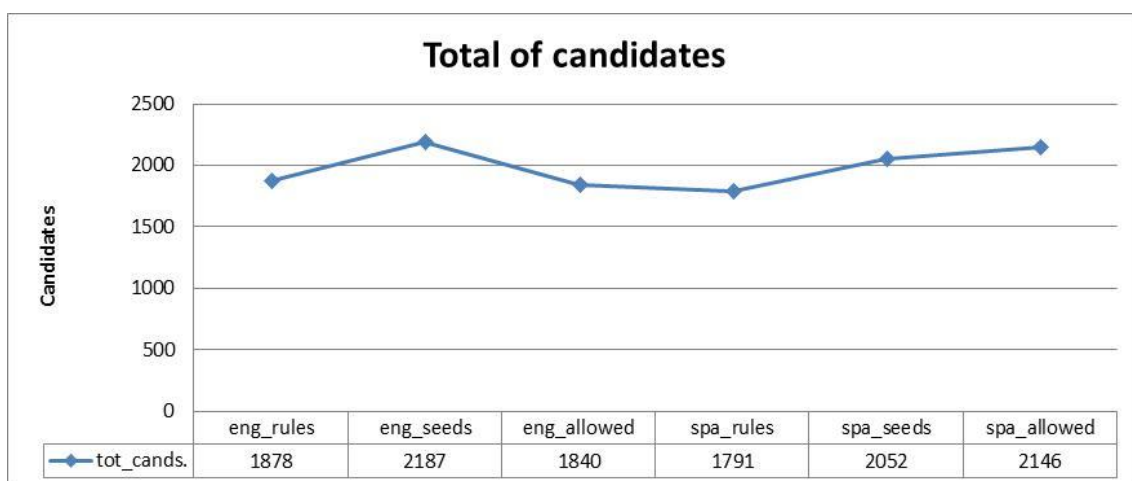


Figure 34. Number of term candidates recognized by each method in in English (left side of the x-axis) and Spanish (right side of the x-axis).

These figures show that the rule-based baseline generates fewer candidates, which affects recall, but increases precision. Its variability is also the closest to 0, which means that the terms retrieved with these rules are always better delimited according to the referent MWTs for both languages.

The recall of seeds and predefined categories has clearly outperformed the baseline, especially in Spanish. The seed approach, a basic lexical method, and the predefined categories approach,

a basic part-of-speech based method, have no statistical difference between them. The fact that these methods do not rely on a specific order of the constituents seems to overcome the problem of unexpected syntactic patterns. As suggested before, these approaches are also expected to generate more noise, that is, less precision given by more irrelevant candidates. It was surprising, however, that the predefined categories generated fewer candidates for English (1,840) than the rules (1,878). This could be due to the use of stop words that could have prevented irrelevant words from becoming candidates even when they could have had an assumed relevant category. In other words, the error of the tagger could have been reduced at some extent. In any case, the outcome of the analyses, including precision and recall, makes the predefined categories method the most effective one in this experiment for MWT recognition. This is worthy to be remarked considering that the adjustments we made to the rule-based method turned it into a very competitive baseline.

It is also worth mentioning that each method also implies a computational cost directly proportional to the needed resources. The rule-based baseline only depends on the tagger and a few rules, the predefined categories approach relies on the tagger and a stop word list, and the seed method needs the tagger, a list of seeds, and a list of stop words. Even when the algorithms can be optimized, the seeds will always imply less efficiency.

This experiment also confirmed the 85% assumption derived from an analysis in the Introduction chapter which suggests that using general noun as seeds, 85% of the MWTs could be retrieved. It was also expected that adding specialized nouns, the 85% could be improved. And indeed, the percentage of recognized MWTs using seeds for both languages not only reached this percentage but also overcame it (eng = 87.15%, spa = 89.82%). Such scores in this evaluation are also remarkable considering that we are using the same seeds used for the initial observations in an automotive engineering corpus.

As a closing remark in this discussion, the configuration of entries in current dictionaries deserves some lines. As it was mentioned before, the test set for this experiment was assembled at random only assuring that the terms came from a reliable source and that, for English, no terms with postmodification were included. Beyond this, no other adjustment was done to the data because we wanted to observe the real effect of our decisions as for our definition of the formal configuration that a MWT should have (i.e., no articles, only one preposition, etc.). And, in fact, some of the non-recognized terms include any of the non-predefined categories or were

wrongly tagged and therefore excluded. However, after reviewing such missed terms, some very atypical cases came out, especially for Spanish:

- ◆ *anualidad que se paga al sobreviviente*
- ◆ *terminal de apertura muy pequeña*

It is comprehensible and predictable that some *ad hoc* resources include even complete sentences, as is the case of some glossaries for software localization that include up to complete multi-sentence messages as entries. But it is not expected a terminological dictionary or database to include entries with the configuration of the ones exemplified above and extracted from the set of missed referent MWTs in our test corpus, and which are difficult to predict.

4.3. Artifact MWT recognition

Keeping in mind that our main goal in this chapter is to align terms and images, we aim now at reducing even more the MWT list obtained in the previous steps. With the assumption that objects in images are more likely to be denoted by concrete nouns, we filter the nominal MWT list so that only nominal multiword terms designating artifacts are left. To do so, we follow two procedures according to the complexity of the document layout, that is, anchor-based selection for simple layouts and noun classification for more complex documents.

4.3.1. Anchor-based selection

If the document is catalog-like, it is usual to find little text around the image, e.g, the artifact's denomination itself. It also seems that naming the image with the vendor's catalog reference code and putting such code next to the artifact's denomination for the user's reference is the rule. In this scenario, which is widely used in online catalogs, the document is parsed and the image name aligned with the matching reference code in the surrounding text.

4.3.2. Noun classification

There is, however, another more complex document layout where the image is surrounded by a greater amount of text. In this case, we perform nominal MWT recognition to obtain a list of MWT candidates (see §4.2 above), and then classify head nouns in order to derive a narrower list of artifact nouns. The resultant artifact noun candidates are aligned with the image.

In order for us to tackle the problem of noun classification, some of the approaches reported in the literature were explored. We describe here the results of three different experiments: 1) a discriminant analysis using the number of images retrieved by concrete and abstract nouns, and the similarity between candidate terms and image names; 2) a Bayesian model based on local linguistic patterns; and 3) semantic classification using lexical senses.

4.3.2.1. Discriminant analysis: retrieved images and edit distance

This first approach to the classification of concrete and abstract nouns was carried out by Burgos & Wanner (2006) with two non-linguistic variables: 1) the number of images that the members of each class (i.e., concrete or abstract) retrieve from the Web using a search engine. The assumption here is that concrete nouns should retrieve a higher number of images than abstract nouns; 2) the similarity between a MWT and an image file name (see Table 16). We assume that if the image file name matches a MWT in the surrounding text, such MWT increases its probability of designating a concrete entity.

For the analysis, 100 concrete nouns and 100 abstract nouns were selected. The concrete nouns were the head of artifact MWTs denoting, for instance, spare parts belonging or related to the automotive engineering domain. The concrete nature of the terms' referents was manually confirmed by means the context. For unknown terms, a dictionary was used. When the complete MWT was not documented in dictionaries, the last modifier at its left was removed and the remaining MWT was searched again. For example, *supercharger drive pulley* was not found as is in the Routledge English Technical Dictionary, but *drive pulley* was with the Spanish equivalent *polea conductora*. The context was enough to determine that *polea* was an artifact. If it were, however, an unknown term, its definition would help clarifying it. For example, *pulley* can be defined as “a wheel-shaped, belt-driven device used to drive engine accessories.” In this example, the word *device* clearly suggests that *pulley* is, indeed, an artifact.

For each of the 200 selected MWTs to be used to query the Web, only one modifier was left. With two lexical items in our terms, we (i) avoid outliers in the values of the retrieved image frequency, (ii) assure a minimum of domain specificity in the image search and (iii) are coherent with the assumed average length of image file names determined in a preliminary analysis. Each MWT was used to retrieve images with a general search engine. For instance, in the case of the MWT *powder-metal connecting rod*, instead of searching for images with the full MWT (which

would lead to the retrieval of 428 images), the search is performed with the shortened MWT *connecting rod* (i.e., the first modifier *powder-metal* is removed). This leads to the retrieval of 2,310,000 images. If we used just the head *rod*, 477,000,000 images are retrieved⁵⁶.

As for the second variable, the Levenshtein edit distance was used to measure the string distance between a MWT and an image file name. The edit distance can be described as the minimum number of steps (substitutions, insertions or deletions of characters and spaces) necessary to convert a word into another. The edit distance is 1 when there are transformations and 0 when no transformations are necessary. To analyze continuous values for this variable, the relative edit distance (RD)⁵⁷ was used to obtain values between 0 and 1.

MWT	Image name	Edit distance
rear axle	rear axle	0
ignition coil	sparky	1
rear axle	stanley rear axle	-0.470588235
throttle valve	throttle	0
oil pan	oilpan	0.166666667
selector lever	image	0.8
cylinder series	series	0

Table 16. Some examples of the relative edit distance measure.

Table 16, shows some examples of the relative edit distance measure for some specific cases. Image file names were cleaned so that underscores or symbols did not interfere in the measurement. If the file name is a substring of the MWT, it is marked as a positive matching; if the file name contains at least one of the MWT's characters, a positive score, although not the lowest, is also given. Each MWT was compared with a maximum of 20 image names and a relative distance mean was established for each MWT.

The tests of equality of group means proved a significant difference between the two measured variables, that is, image frequency and relative edit distance. 74.4% of originally grouped cases were correctly classified using these variables in a discriminant analysis. A detailed analysis of the results shows, however, that there is bigger variance within the values of concrete nouns than within abstract nouns. This variance could be due to the fact that sometimes concrete

⁵⁶ Number of retrieved images updated using Google on 7/16/2013.

⁵⁷ RD = number of transformation steps / possible maximum transformations.

nouns retrieved very general, irrelevant images and some abstract nouns also retrieved an unexpected high number of images.

The ways image files are named continue to be relevant for our purposes, as will be shown below. However, the capacity of concrete and abstract nouns to retrieve images seems to be affected by external factors such as the Web structure and continuous variability. This is the reason why we approached more linguistic variables for noun classification which are described below.

4.3.2.2. Bayesian classifier based on local linguistic features

In the natural language processing tradition, the most successful systems of lexical classification (normally referred to as *lexical acquisition*) are based on the linguistic idea that the contexts where words occur are associated to particular lexical types. Although the methods are different, most of the systems work upon the syntactic information on words as collected from a corpus and develop different techniques to decide whether this information is relevant or whether it is noise, especially when there are just a few random examples of each class. For this purpose, we have used a Bayesian model of inductive learning based on Bel et al. (2008) to classify concrete and abstract nouns.

Given a hypothesis space (that is, all what a word can be, according to a typology) and one or more examples of a noun to be classified, an automatic learner evaluates all hypotheses for candidate noun classes by computing their posterior probabilities, proportional to the product of prior probabilities and likelihood. We have produced a probabilistic version of a lexical typology and we have used it as a representation of the lexical knowledge of a language.

In order to obtain the likelihood, the information of the class is related to the expected contexts where the members of a class might appear. Such a characterization is done by means of syntactic features that describe distributional classes. Syntactic features can be understood as a linguistic function made up by an ordered set of words and/or morpho-syntactic tags; likewise, morphological features can be understood as a linguistic function made up by an ordered set of units belonging to the internal structure of words. See the examples below as an illustration:

Syntactic or morphological feature	Concrete	Abstract
Short epithet + <i>noun</i> e.g., difícil <i>reto</i>	No	Yes
Prefix <i>des-</i> e.g., <i>desprestigio</i>	No	Yes
Suffix -ión, -ncia, -ento, -ismo	No	Yes
Noun + prepositional phrase e.g., respuesta a antígeno	No	Yes
Suffix -tor(a), -dor(a)	Yes	No

Table 17. Example of lexical typology.

The class of a particular noun is determined by averaging the predictions of all hypothesis weighted by their posterior probabilities. More technically, for each syntactic feature $\{sf_1, sf_2, \dots, sf_n\}$ of the set SF represented in the lexical typology of reference (see Table 17), we define the goal of our system to be the assignment of a value $\{no, yes\}$ that maximizes the result of a function $f: \sigma \rightarrow SF$, where σ is the collection of occurrences of a noun that we call signature, where each occurrence is a vector. To assign the value, every occurrence of the noun is considered as a cumulative evidence in favor or against of having each syntactic feature. Thus, the function $Z'(SF, \sigma)$, shown in (1), will assess how much relevant information is got from all the vectors given every syntactic feature sf_i and each value SF_x , and a particular word signature σ containing \varkappa different vectors, $\sigma = \{v_1, v_2, \dots, v_\varkappa\}$. A further function (4) will decide on the maximal value in order to assign $sf_{i,x}$.

$$(1) Z'(sf_{i,x}, \sigma) = \sum_j^{\varkappa} P(sf_{i,x} | v_j)$$

$P(sf_{i,x} | v_j)$ is assessed in (2) with the application of Bayes Rule for solving the estimation of the probability of a vector conditioned to a particular feature and value.

$$(2) P(sf_{i,x} | v_j) = \frac{P(v_j | sf_{i,x})P(sf_{i,x})}{\sum_k P(v_j | sf_{i,k})P(sf_{i,k})}$$

For solving (2), we assume that the prior $P(sf_{i,x})$ is computed on the basis of the typology.

For computing the likelihood $P(v_j | sf_{i,x})$, as each vector is made of m components, that is the linguistic cues⁵⁸ $v_{\zeta} = \{lc_1, lc_2, \dots, lc_m\}$, we proceed as in (3) on the basis of $P(lc_l | sf_{i,x})$ which is the likelihood of finding the noun in a particular context given a particular syntactic feature. The likelihood for each lc is assessed given every syntactic feature from the lexical typology.

$$(3) P(v_j | sf_{i,x}) = \prod_{l=1}^m P(lc_l | sf_{i,x})$$

Finally, Z as in (4), is the function that assigns the syntactic features to the noun signatures⁵⁹.

$$(4) Z = \left\{ \begin{array}{l} Z'(sf_{i,x = yes} | \sigma) > Z'(sf_{i,x = no} | \sigma) \rightarrow yes \\ Z'(sf_{i,x = no} | \sigma) > Z'(sf_{i,x = yes} | \sigma) \rightarrow no \end{array} \right\}$$

Experimental results

For the experiments, we used a Spanish part-of-speech tagged automotive engineering corpus made up of 12 texts and 17,278 words. A data subset of 100 abstract nouns and 100 concrete nouns was manually selected and annotated.

Table 18 shows the validation results where a class was assigned to each noun in the test corpus. The evaluation was carried out with the F -measure. 0.74 in the F score shows that the approach as well as the linguistic cues selected to induce the lexical classes reasonably fit the problem.

During the experimentation, it was clear that one of the biggest problems for lexical classification is the lack of local information. We did not find discriminant linguistic cues for a good number of noun occurrences. For abstract nouns, a greater number of linguistic cues were defined than for concrete nouns. Such difference is linguistically motivated and led classifier to assign a class to those nouns whose profiles did not present any linguistic cue based on the information learned from the signatures.

⁵⁸ A linguistic cue is the result $\{0,1\}$ of verifying with a regular expression if the noun context presents a morphological or syntactic pattern characterized by tags, words and the position of the relevant noun.

⁵⁹ In the theoretical case of having the same probability for *yes* and for *no*, Z is undefined.

	Precision	Recall	F-measure
Abstract	0.66	0.95	0.78
Concrete	0.93	0.55	0.69
Totals	0.74	0.74	0.74

Table 18. Evaluation results.

4.3.2.3. Semantic classification using lexical senses

The lack of morphological or syntactic cues in the occurrence of certain nouns exhausts the possibilities of a classification method based on local features, as shown above. This suggests the need of higher level information to try to improve noun classification, that is, semantic information. We consider that noun classification can be seen as a word sense disambiguation (WSD) problem and that it can be approached using lexical semantics. This is the reason why we also carried out some experiments to assign WordNet (Fellbaum, 1998) supersenses to a set of English and Spanish nouns. For the annotation, the SuperSenseTagger (SST, Ciaramita and Altun, 2006) and Freeling (Carreras et al., 2004) were used.

The SST, currently available for English and Italian, is a Hidden Markov Model (HMM)-based tagger which uses a probabilistic model to determine hidden values (classes) from observations (nouns). In practice, this means that the SST does not assign the most frequent sense of a word as other applications do. Instead, it calculates the probabilities by means of a HMM to assign the most appropriate sense according to the context. This explains the fact that different occurrences of the same word could be annotated with different senses in distinct contexts. For example, *valve* sometimes is assigned the sense *body*, but other times it receives the sense *artifact*.

In a previous experiment (Burgos, 2009), the SST was used to classify the same noun subset (100 abstract and 100 concrete) that was used in §3.3.2.1. The input for the SST is a set of nouns part-of-speech tagged with the TreeTagger (Schmid, 1994), and the output is a set of nouns with WordNet supersense annotation. Given the domain of the observations of this work, that is automotive engineering, the relevant nouns are those annotated as *artifacts*. See Table 19 some examples of nouns annotated with the SST:

Concrete	Abstract
converter NN I-noun.artifact	zone NN B-noun.location
stack VV B-noun.artifact	segments NNS B-noun.artifact*
blades NNS I-noun.artifact	sections NNS I-noun.location
coupling NN B-noun.artifact	displacement NN B-noun.act
bearing NN I-noun.artifact	mixing VVG B-noun.act
nozzle NN B-noun.artifact	periphery NN B-noun.shape
coils NNS I-noun.artifact	series NN B-noun.group
clutch NN B-noun.act*	headroom NN B-noun.quantity
plug NN I-noun.artifact	objectives NNS B-noun.cognition
covers VVZ B-noun.artifact	teams NNS B-noun.group
pan VV B-verb.motion*	injection NN B-noun.act

Table 19. Nouns annotated with WordNet supersenses.

This example shows some cases marked with asterisks that, according to the context, received a wrong sense. The performance of the SST shown in Table 20 is encouraging, though:

	Concrete	Abstract	Without annotation	Not analyzed
Concrete	81	14	1	4
Abstract	8	90	0	2

Table 20. Results of the semantic annotation.

These results show that out of 95 concrete nouns, 81 were correctly annotated, and that out of 98 abstract nouns, 90 were assigned the right sense. The overall precision was 0.855.

In spite of these encouraging results, the small sample, and the fact that the SST is not available for Spanish motivated further experiments with semantic annotation. For English, we used the SST again with a bigger sample of 1,876 MWT candidates. For Spanish, the UKB algorithm (Agirre and Soroa, 2009) integrated in Freeling was used to annotate 2,138 MWT candidates with the most probable sense.

Likewise, the same samples in both languages were annotated with the most frequent sense (MFS) as baseline using Freeling. Any nouns missed by the tagger were manually annotated

using WordNet 3.0 and EuroWordNet⁶⁰. The most frequent sense is considered a good baseline with an acceptable performance that is not easy to beat (cf. McCarthy et al., 2004). Precision and Recall were used for assessment in both languages for both the most probable and the most frequent sense. The results of the evaluation are illustrated in Figure 35 and discussed below.

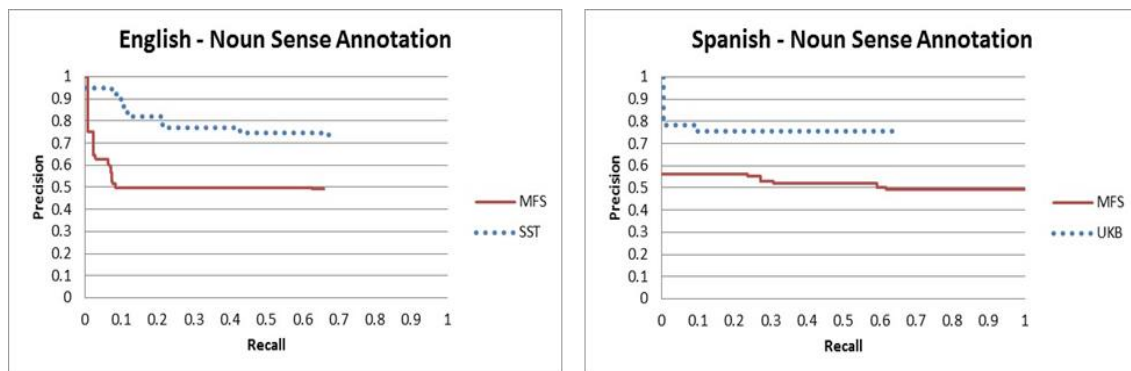


Figure 35. Precision and recall for artifact noun annotation with SuperSense Tagger (SST), UKB algorithm, and the most frequent sense (MFS).

	MFS (baseline)			SST			UKB		
	P	R	F1	P	R	F1	P	R	F1
English	0.49	0.66	0.56	0.74	0.68	0.71	--	--	--
Spanish	0.49	1	0.66	--	--	--	0.76	0.65	0.70

Table 21. Precision and Recall for three lexical sense-based methods of artifact noun annotation, including the baseline.

Table 21 shows that the SST keeps performing better than the MFS for English. Our analyses confirm the trends shown in Ciaramita and Altun (2006). The MFS remains a strong baseline, but the SST reports better accuracy. The UKB also outperforms the MFS. We managed to reverse the trends reported by Agirre and Soroa (2009), who obtained better results with the MFS than with the UKB. Higher scores for both languages are still desirable, but it seems the performance of the UKB and the SST was affected by the characteristics of the data set we used here. As for the Spanish test set, for example, we used a small corpus made up of contexts from the telecommunications and assurance subject fields. While the telecommunications

⁶⁰ EuroWordNet was consulted using the Multilingual Central Repository at adimen.si.edu.es/cgi-bin/wei/public/wei.consult.perl.

domain certainly has more artifact referents, the assurance domain uses more abstract referents that make the search space noisier and that could affect the final recognition outcome. It is also interesting to note that decisions for noun sense annotation by both the SST and UKB are mostly made based on the neighbor words around the candidate noun. Thus, when we ran a first classification experiment on Spanish terms with no additional context, an F score of 0.58 was reached. However, when a three-word context was added to the left of each candidate in a second experiment, the F score boosted to 0.7.

It is expected, then, that the more specialized the text, the better the performance of the MWT recognition task. Likewise, the domain plays an important role for artifact MWT recognition. As suggested in the assumptions presented in the Introduction chapter (§1.3.1) the occurrence of terms and artifact nouns, as well as of related images, is a language- and domain-independent phenomenon, although it can be higher in some languages and in some domains. We also expect that the application of the methods described here for artifact MWT recognition on appropriate domains such as automotive online catalogs boosts the overall performance of a real world application, as describe in the next chapter.

5. IMPLEMENTATION OF THE BC-MODEL

5.1. Introduction

The previous chapters described and characterized contexts, requirements, and instances to give theoretical and practical support to the bimodal co-occurrence (BC) model that was introduced in Chapter 1 (§1.2). Thus, while Chapter 3 explored the features of the selected search space for this proposal, Chapter 4 presented and evaluated the methods and techniques necessary for the application of the textual component of the model. This theoretical and experimental background enabled the implementation of a functional software prototype of the BC-model that is described and evaluated in the present chapter. The first part of the chapter contains a simple software design description (SDD) of the prototype and the second part provides an evaluation of the system's performance focusing on the CBIR component. The SDD uses some guidelines of the IEEE Standard for Information Technology —Systems Design— Software Design Descriptions (1016-2009) to summarize the system design. It presents 1) a behavioral model which includes use cases and an activity model, 2) a structural or architectural model with its component and deployment models, and 3) a class diagram. The user interface is also described and explained, which also serves as a basic user manual for the prototype.

5.2. System overview

The software has been named *BC-Trans* (Bimodal Co-occurrence-Based Translation Software). It either finds translations for artifact multiword terms (MWTs) or finds terms for artifacts in photographic images. User can type a term and get translations and images of his/her term or can upload a photographic image of an artifact and get the term for the artifact. Terms and images are stored in a server-based database. User accesses BC-Trans and gets results via web browser⁶¹. The system was designed on the basis of the BC Model, which means that it expects to be queried with artifact terms and/or photographic images.

⁶¹ The prototype can be accessed at <http://grupotnt.udesa.edu.co/bc-trans>.

Audience. The software is intended to be used by translators and linguists. Even other users in a commercial setting could find it useful when there is the need of getting terms in a foreign language that enable them to explore the existence of spare parts in foreign markets.

Programming languages. For the integration and implementation of most of the modules, Python programming language was mainly used. The web interface was implemented using HTML and Jinja2 templating language. The image search engine DORIS was developed using Java.

Interface language and language pairs. The graphical user interface language is English. Available languages for user query translation are English and Spanish.

Paper prototype. Prior to the prototype design, a series of tests took place with the purpose of tuning up the methods in the backend and the user interface. For the tests, an experiment with a *paper* prototype was designed. This paper prototype consisted of a simulated version of the user interface. It was created in PDF format. 6 users with different profiles were assigned two tasks and asked to follow the instructions in the simulated screens. They were observed during the experiment and their questions and decisions were recorded. The two assigned tasks correspond to the two possible use cases:

- ❖ Taks 1: Use this service to find an English translation for the Spanish term “*regulador de voltaje*”
- ❖ Task 2: User this service to find an English translation for the Spanish term “*estabilizador de corriente*”

Task 1 emulated a scenario where the query matches a term in the system’s database, while in Taks 2 there is no match for the query so the user is prompted to upload an image. The complete paper prototype is attached in the Annex.

5.3. System Design

The subsections below present the system’s behavioral model and as well as the structural/architectural model.

5.3.1. Behavioral Model

5.3.1.1. Use case

A use case is a specific way of using the system by utilizing some part of its functionality (Jacobson, 1992) and describes the way a system is used by its actors to achieve their goals (Armour and Miller, 2001). BC-Trans foresees two cases: 1) find a translation for a term, and 2) find a term for an artifact in an image. A description of the use cases is illustrated in the activity model below.

5.3.1.2. Activity model

Figure 36, shows the activity diagram. It shows how a query, i.e. a term, in source language is submitted by the user for translation. The query is normalized and so it is any term in the bilingual database. Similarity is computed between the query and the terms in the database. If the query is a MWT, to compute this similarity, we start with the head noun (HN), then with the first modifier, then the second modifier, and so on. When computing this similarity, we give most of the weight to a match between head nouns (HNs) and less weight to matching modifiers or modifiers matching HNs. Then, if two terms match in their HN, the next step is to verify if the first modifiers also match, and so on. This way, scores are assigned and a ranking defined before returning the most relevant terms to the user.

A list of similar or related terms and associated images, still in source language, is returned to the user. The user selects and submits relevant results. Then, the system returns a list of proposed candidate translations and associated images. The user selects and submits relevant translations and the system returns a bilingual term record with terms and images in both languages.

If the query does not match any term in the database or the user is not satisfied with the returned list of related terms in the source language, the user is prompted to upload an image of the artifact whose term is to be translated. Image matching is carried out by DORIS between the user's image and the image database. Then, the system returns a list of matched images and associated terms in target language. The user selects and submits relevant results and the system returns a bilingual term record with terms and images in both languages.

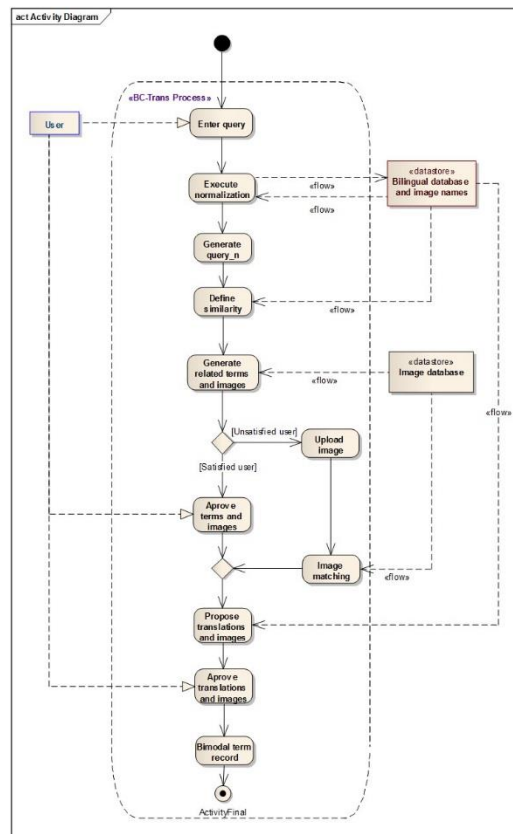


Figure 36. Activity diagram.

5.3.2. Structural/Architectural Model:

5.3.2.1. Component model and class diagram

Figures 37 and 38 depict the major components of the system and the class diagram respectively. They are also briefly described below.

- ❖ Related terms search: Given a user's term, it returns a list of 4 rows in *csv* format which has related terms in source language.
 - Term normalization: It returns a term after normalization. Normalization includes lowercasing, lemmatization and temporary removal of stopwords.
 - Head noun modifiers identification: In order to make matching easier, it takes a MWT and returns it in reverse order (for English) so we have the head noun first and then the modifiers.

- Similar term finding: It matches the normalized user's term with normalized rows in dictionaries. It returns a list of rows for those matches on head nouns and modifiers in descending order (i.e., highest ranking first).
- ❖ Dictionary processing: It loads data from the database and returns a list of proposed translations according to source language relevant terms selected by the user.
- ❖ Image matching: It parses data from DORIS to get image names and returns images.
- ❖ Similar image finding: It passes image file to DORIS which analyzes the image features and searches the database for similar images depending upon target language. It processes output from DORIS to get the most similar images and returns a list of filenames of similar images.
 - Image path generation: It returns the location of the directory containing images for a term.
 - Image name and info generation: It returns information for an image filename.
 - Related images generation: Given an image filename, return the number of similar images with associated terms and URLs.

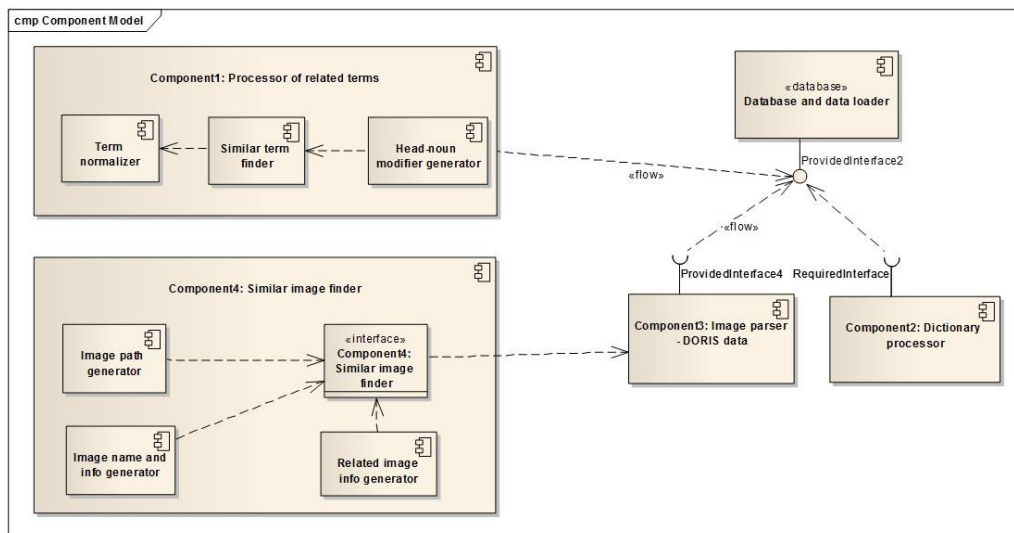


Figure 37. Component model

5.3.2.2. Deployment model

The system basically features client-server architecture. The text and image applications which process user's queries are hosted in a server. Data is stored to and retrieved from databases. Some of the user's decisions are also stored. A graphical user interface is used to send queries and to visualize results. Figure 39 illustrates the system's deployment model.

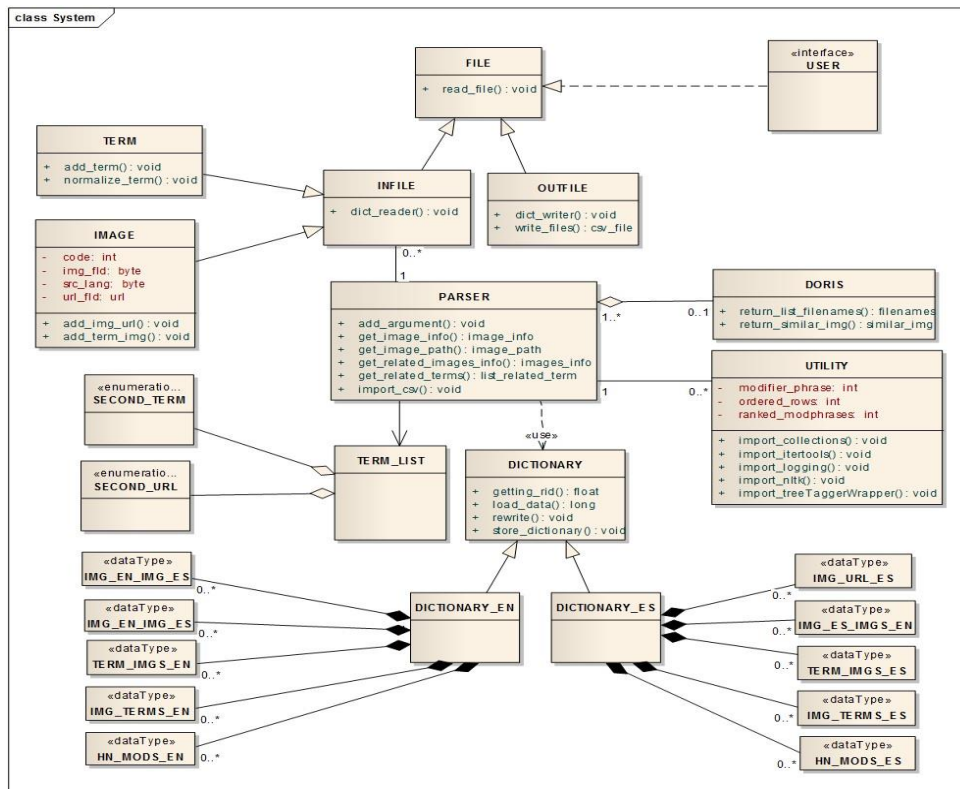


Figure 38. Class diagram.

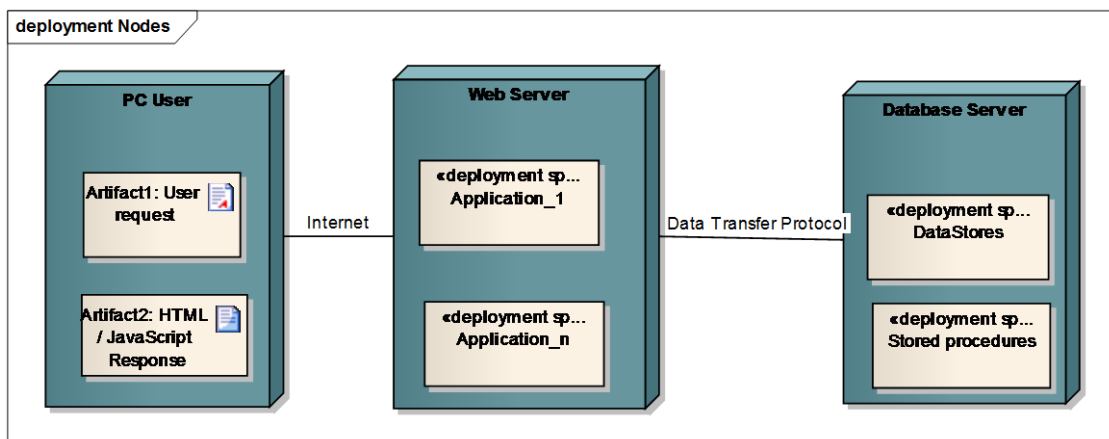


Figure 39. Deployment Nodes.

5.4. User interface

A description of the graphical user interface is presented below. It can also serve as a basic user manual for the software prototype.



The screenshot shows a form titled "Choose your language pair: *". It contains two radio buttons: "English -> Spanish" (selected) and "Spanish -> English". Below this is a text input field labeled "Term to translate: *". At the bottom left is a "Submit" button.

Figure 40. User is prompted to enter query.

Query submission:

In this screenshot (Figure 40), the user is prompted to select the language pair and to type the term in the *Term to translate* field. Both the value of the language pair radio buttons and the text field are mandatory. After selecting a radio button and entering the term, the query is submitted by clicking on the *Submit* button.

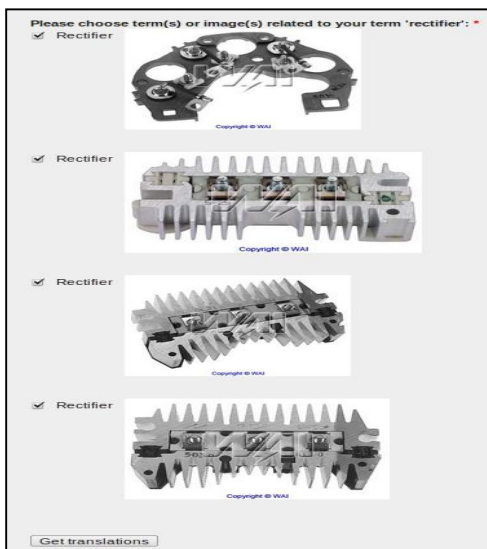


Figure 41. System returns related terms.



Figure 42. System returns candidate translations.

Select related terms in source language:

This screenshot (Figure 41) presents the returned related terms and associated images in source language for the query **rectifier**. Normalization has already taken place in the back end, but terms are returned as they are in the database with forms, capitals and functions words. For this example, there are source language exact matches in the database for the query **rectifier**. Results deemed relevant by the user are selected and submitted by clicking on the *Get translations* button. Checking at least one of the check boxes is mandatory.

Select candidate translations:

In this screenshot (Figure 42), the system returns a list of candidate translations and associated images. The user selects and submits relevant translations by clicking on the *Get translations* button. Checking at least one of the check boxes is mandatory.

Results shown in this example are relevant since the *a priori* expected translation in Spanish for **rectifier** (**rectificador**) is not among the results. Instead, the user makes his/her decision based on the visual representation and selects the Spanish equivalent **portadiodo chevrolet valeo 100-120A** which has associated an image of a rectifier.

It is certainly user's general and specialized knowledge that helps him/her know that **portadiodo** and **rectifier** are equivalents and that 1) the modifiers of the Spanish term (**chevrolet valeo 100-120A**) are external ones not inherent to the artifact itself, and 2) that the functions words (*para, de*) in the target term have been omitted as a characteristic of the search space (see Chapter 3, §3.4.2).

Term record:

This screenshot (Figure 43) presents the term record with the final results of the query. The record contains one entry in source language and at least one entry in target language. Source language entry contains ISO language code, term and term source. Target language entry contains ISO language code, term, term source and associated image.

Upload image:

Screenshot in Figure 44 prompts user to upload an image when there is no match for the query in the bilingual database. In the example, there was no match for the Spanish term **rectificador**.

Uploading an image is mandatory to continue the processing of the query and is carried out following standard procedures of the operating system.

Select an image:

This is the operating system standard window where the user is prompted to select the image of the artifact (Figure 45). Once the image is uploaded, image matching will be carried out by DORIS in the back end. The results of this operation are shown in the next screenshot.



Figure 43. System returns term record

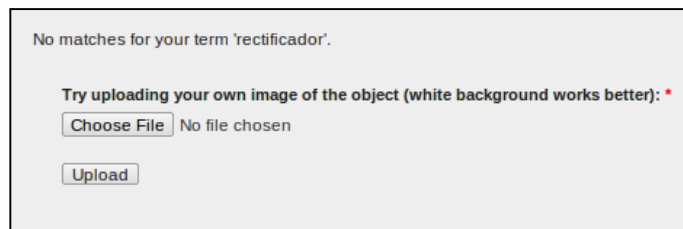


Figure 44. User is prompted to upload an image.

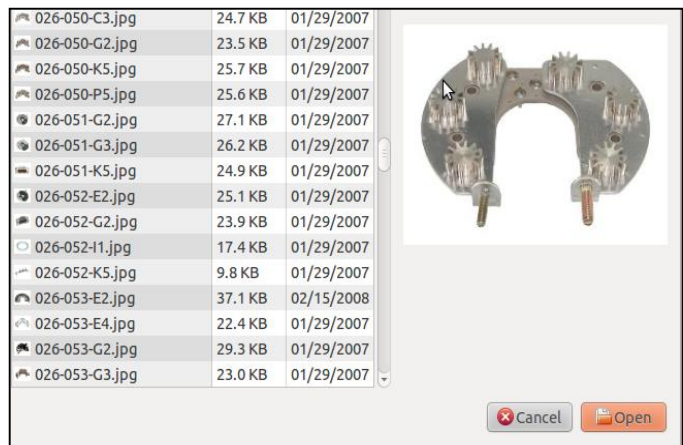


Figure 45. User selects an image to upload.



Figure 46. Select candidate target terms.



Figure 47. Term record.

Select candidate target terms:

In this screenshot (Figure 46) the system returns a list of matched images and associated target terms. The user selects and submits relevant translations by clicking on the *Get translations* button. Checking at least one of the check boxes is mandatory. In this example, the user deems relevant three of the translations (*rectifier*) for his/her term *rectificador*.

Term record:

This screenshot (Figure 47) presents the term record with the final results of the query. The record contains one entry in source language and at least one entry in target language. Entries in source and target language contain ISO language code, term, term source and associated image.

5.5. System evaluation

After the description of the prototype, it can be seen in a practical application how the visual representation, i.e., the image, plays a pivotal role in the BC model. Its role is crucial since the link between source and target images is expected to bridge the gap between source and target terms too. For the purpose of this research, such a link between source and target images is to be established by means of a content-based image retrieval (CBIR) application. This means, therefore, that the performance of the CBIR application reflects, to a great extent, the performance of the whole prototype. The sum of accuracy of the CBIR component plus the precision of the artifact term recognition techniques presented in Chapter 4 is to be read as the effectiveness of the practical application of the BC model, since matching two images also means matching two equivalent terms. It is upon this premise that we devote this section to the evaluation of DORIS (Domain-ORiented Image Searcher), a CBIR software designed according to the characteristics of the images found in our search space and which are described later in this chapter.

5.5.1. Evaluation of the CBIR component

Although technical details about DORIS's algorithms can be found in Jaramillo and Branch (2009), it is worth mentioning here that the system uses MPEG-7-based shape descriptors. The authors implemented Zernike (Teague, 1980) moments which seem to adequately fit the characteristics of our prototypical image. Thus, while SIFT and SURF would be more appropriate for colorful and general images, MPEG-7 descriptors used by DORIS better represent the monochromatic, noiseless, and specific-domain nature of our images. By combining Zernike moments, mass, and eccentricity, the authors report 90% of precision using empirically set parameters.

Besides the accuracy reported by its authors, it was deemed necessary to test DORIS' performance with unseen images directly taken from our search space in order to be able to predict effectiveness for artifact term translation. In order to evaluate DORIS' performance, three sets of images labeled with their artifact terms were selected: two sets (A and B) from a different website in Spanish each, and another set (C) from a single site in English. All of the sites belong to the automotive engineering category defined in Chapter 3 (§3.2.2). Set A has 401 images, set B has 6,849 images and set C has 11,041 images. The evaluation was carried

out in terms of the percentage of relevant images retrieved as well as of the mean average precision (MAP). Three evaluation tasks were carried out with each set.

It is worth noting that even though it might seem that we are addressing only image matching evaluation in this subsection, we are at the same time testing image-based artifact term translation. This can be especially observed with the full term matching (FTM) and head noun matching (HNM) tasks below:

1. **Image name matching (INM):** In this task, every single image of each set was used as example to match all the images within the same set. Matching is verified using image file names. The assumption here is that an image i should retrieve from its own set a subset of similar images, being i itself the first in the ranking or at least being it well-ranked. However, in case i is not well-ranked or not retrieved at all, it cannot be said that the matching task failed because other images with different file names could represent the same artifact of i ; that is why a second task described below was also performed.
2. **Full term matching (FTM):** In this task, every single image of each set was used as example to match all the images within the same set. The assumption here is that an image i labeled with an index term t should retrieve from its own set a subset of similar images also labeled with t , that is, images of the same artifact. Matching is verified using image index terms. However, in case that some or all of the retrieved images are not labeled with t , it cannot be said that the matching task has poorly performed because t could be a MWT whose head noun (HN) could happen to match other HNs of the retrieved image labels. In case of positive HN matching, the retrieved images can continue to be relevant given the specificity of the application domain, i.e., automotive engineering parts and accessories; in other words, in this scenario the retrieved images might not represent exactly the same artifact of i , but a variation of the artifact; that is why the third test below was also performed.
3. **Head noun matching (HNM):** In this task, every single image of each set was used as example to match all the images within the same set. The assumption here is that an image i labeled with an index term t and a head noun t_h , should retrieve from its own set a subset of similar images labeled with a term whose head noun is equal to t_h ; that

is, images with variations of the same artifact. Matching is verified using head nouns of image index terms.

5.5.2. Evaluation results

Let us now present some figures derived from the experiments described above using DORIS for image matching in the frame of our BC model. It is worth to say that only the first three results in each task will be considered for this evaluation. The decision on the number of results for evaluation was motivated by the fact that the prototype described in this chapter was initially designed to return the first three results to the user. Therefore, for the sake of usability and interaction with the application of the model, it is important to foresee how the information was going to be presented to the end user and three results seemed to be a reasonable number.

5.5.2.1. The INM task

For each image i , the INM task generated both i 's ranking within the retrieved images (from 1 to 10) as well as its distance of similarity to itself in a scale which ranges from 0.0 to 1. Theoretically, 1s and 0s respectively would be expected, but given the variability and the size of the samples, it is not always the case, as shown below.

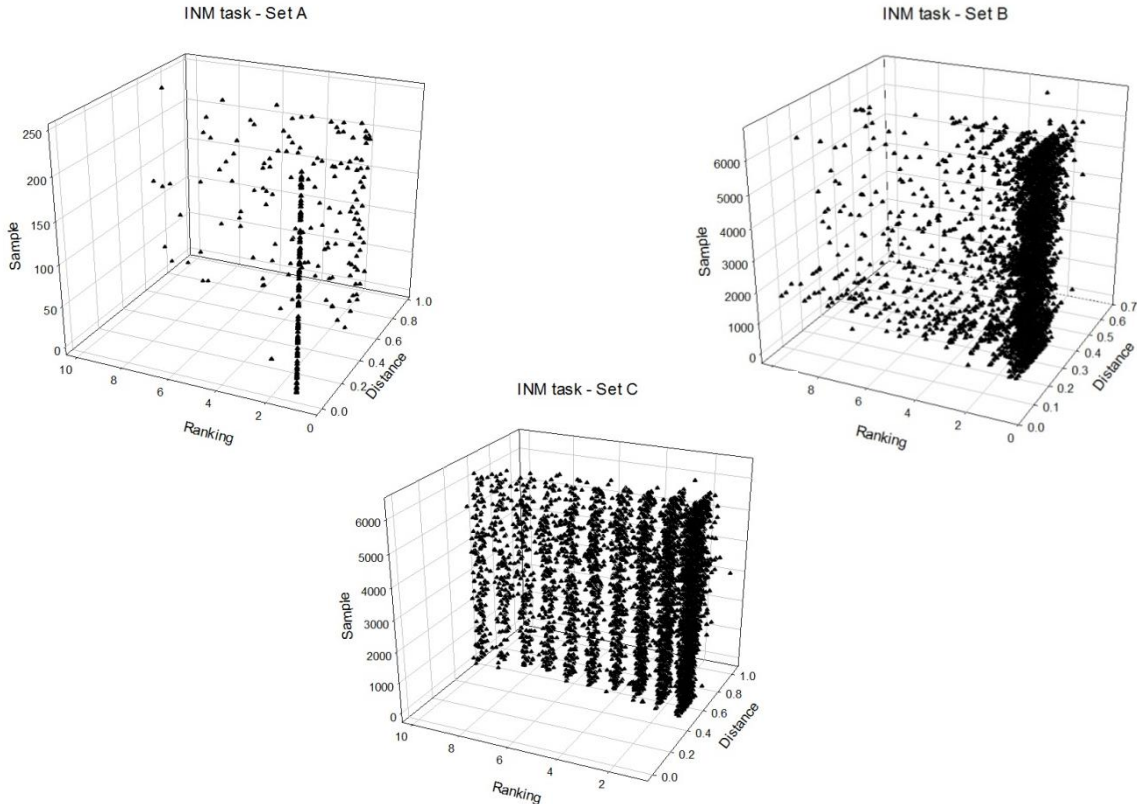


Figure 48. Ranking and distance for image sets A, B and C in the image name matching task.

Figure 48 illustrates the results of the image name matching (INM) task for the three sets. The Figure shows the images ranked within the first ten positions after the task. Thus, 244 images were ranked out of 401 images in set A (60.84%), 6,604 out of 6,849 in set B (96.42%) and 6,302 out of 11,041 (57.07%) in set C. The graphics also show a tridimensional space with the size of each sample after the INM task as well as the rank and the distance of every image i with respect to i itself. As for the image ranking, and out of the total of images for each set, the matching operations ranked i within the first three positions 45.38% of the times in set A, 91.31% in set B and 41.4% in set C.

As for the similarity distance, in set A there is a subset (37.7%) of images with a distance 0.0 but then most of the other images (60.6%) jump to be within a distance range between 0.5 and 0.76. Set B, on the contrary, has a 99.83% of its images within a distance range between 0.25 and 0.5, although none of them at 0.0. In set C, most of the images (94.52%) are grouped within a distance that ranges from 0.45 to 0.75.

5.5.2.2. The FTM task

The full term matching (FTM) task generated for each image i labeled with its index term t the three most similar images with their respective index terms as labels. Then, t was matched to each label of the retrieved images. Positive matches were marked with 1 and negative matches were marked with 0. The matches for each image are summed, so the graphics represent the images with 0, 1, 2 and 3 full term matches.

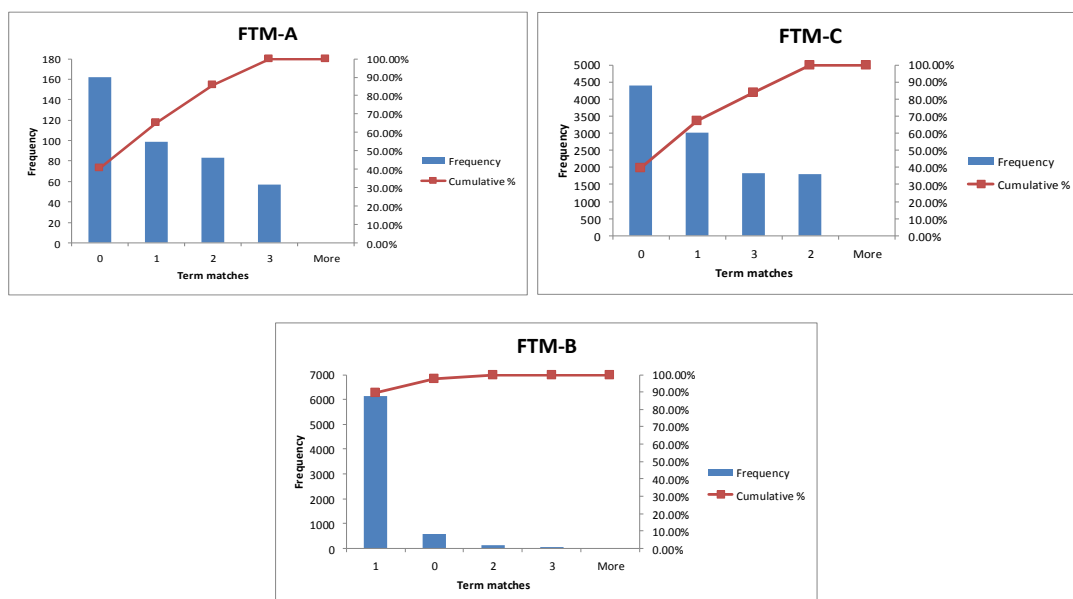


Figure 49. Evaluation results of the FTM task for the three image sets.

Figure 49 shows the evaluation results of the FTM task for our three sets. For set A, there was 59.6% of full term matches; for set B, there was 91.58% and for set C there was 60.27%.

5.5.2.3. The HNM task

The head noun matching (HNM) task generated for each image i labeled with its index term t the three most similar images with their respective index terms as labels. Then, the head noun t_b of t was matched to each head noun in the label of the corresponding retrieved images for i . Positive matches in head nouns were marked with 1 and negative matches were marked with 0. The matches for each image are summed, so the graphics represent the images with 0, 1, 2 and 3 head noun matches.

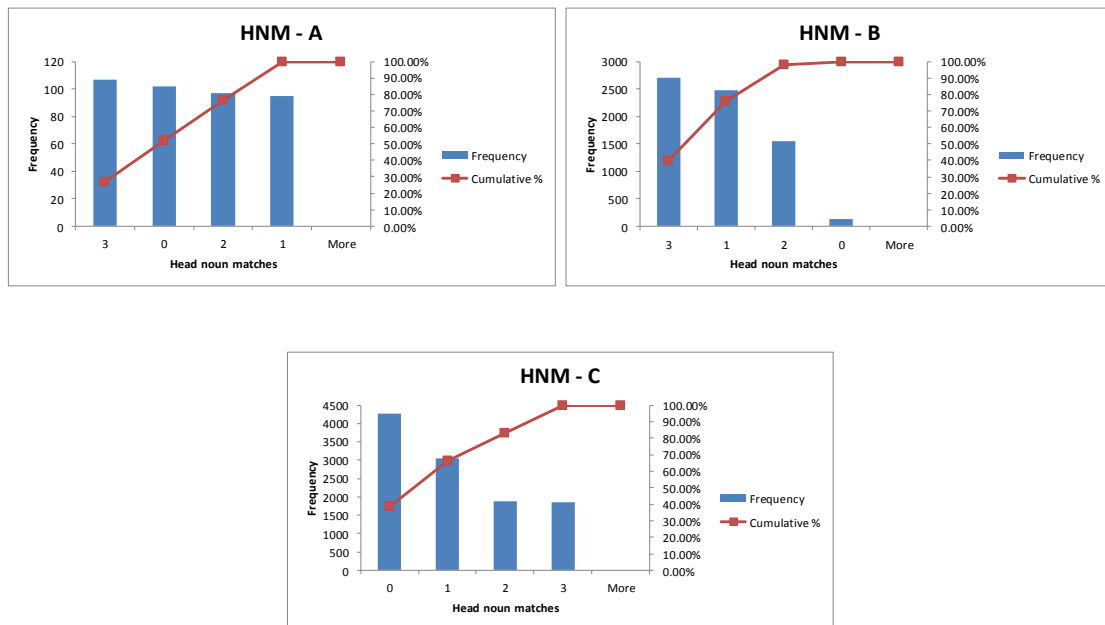


Figure 50. Evaluation results of the HNM task for the three image sets.

Figure 50 shows the evaluation results of the HNM task for our three sets. For set A, there was 74.56% of head noun matches; for set B, there was 98.2% and for set C there was 61.28%.

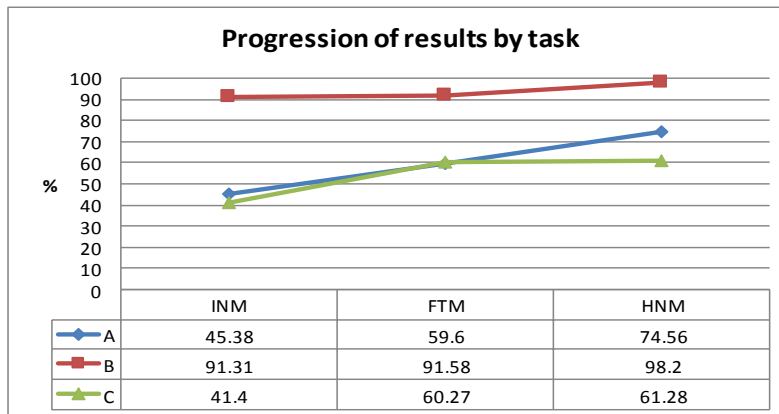


Figure 51. Progression of results by task for each image set (A, B, and C).

Figure 51 summarizes the progression of results by task for each set. All of the sets underwent an improvement in the percentage of matches. It is remarkable, however, the excellent performance of DORIS with set B from the first task and how it improved in a 6.89% through the other two tasks to reach an encouraging **98.2%**.

5.5.2.4. Mean average precision

The mean average precision (MAP) is a widely used measure for information retrieval. Here, MAP is also used in order to be able to compare DORIS' performance with the effectiveness of other similar approaches reported in the literature, especially those presented in different editions of the ImageCLEF track.

Table 22 presents the mean average precision for each set as well as the number of queries for each set. Evaluation for MAP scores was also for the first three retrieved images.

Set	Queries	MAP
A	401	0.465
B	6849	0.655
C	11041	0.332

Table 22. MAP scores for the head noun matching task.

Compared to the MAP scores reported by other researchers (see Tables 23 and 24), DORIS' scores are encouraging and superior. Even the lowest score in our tests (0.332 for set C) is

somewhat close to the highest scores in the ImageCLEF tasks and is higher than the highest score in the visual modality.

Task	Modality	Best MAP
Medical image retrieval task (Müller et al., 2007)	Textual	0.3962
Wikipedia image retrieval task (Tsirikika et al., 2011)	Mixed	0.3880
Photographic retrieval task (Clough et al., 2006)	Mixed	0.385
Photographic retrieval task (Grubinger et al., 2007)	Mixed	0.3175
Medical image retrieval task (Müller et al., 2008)	Mixed/Text	0.29
Wikipedia image retrieval task (Popescu et al., 2010)	Mixed	0.2765
Wikipedia image retrieval task (Tsirikika et al., 2009)	Textual	0.2397

Table 23. Best MAP scores for some ImageCLEF tasks - Different modalities.

Task	Modality	Best MAP
Medical image retrieval task (Müller et al., 2007)	Visual	0.2328
Photographic retrieval task (Grubinger et al., 2007)	Visual	0.1890
Photographic retrieval task (Clough et al., 2006)	Visual	0.1010
Wikipedia image retrieval task (Popescu et al., 2010)	Visual	0.0553
Medical image retrieval task (Müller et al., 2008)	Visual	0.04
Medical image retrieval task (Müller et al., 2010)	Visual	0.0358
Wikipedia image retrieval task (Tsirikika et al., 2009)	Visual	0.0079
Wikipedia image retrieval task (Tsirikika et al., 2011)	Visual	0.0044

Table 24. Best MAP scores for some ImageCLEF tasks - Visual modality

There are certainly some aspects that play a role in these results and that have been mentioned before. The Medical Image Retrieval Task mainly deals with images generated by imaging technologies such as x-ray radiography, ultrasound and computed tomography. Features in this kind of images are fuzzier and more difficult to identify. Likewise, the Wikipedia Image Retrieval Task as well as the Photographic Retrieval Task face additional challenges posed by noisy images.

The type of images associated to the BC model, on the contrary, do not have noisy backgrounds and present defined characteristics that allow for the identification and extraction of strong features. In the results from the analysis above there is a set of images which clearly maximizes CBIR performance compared to the other two sets and to the MAP scores reported by the literature, that is, set A. Such encouraging results suggest that the features of images in set A, as a whole, could constitute the prototypical image for an optimal performance of the BC model. The section below proposes a definition for the BC model prototypical image based on image features and visual perception.

5.5.2.5. The prototypical image

Here, we try to give a description of the characteristics of the prototypical image from the visual perception, on one hand. On the other hand, although image analysis is out of the scope of the present research, a statistical comparison of prominent features in the three sets of images is also presented. The expectation here is to identify the features in set A that define our prototypical image.

5.5.2.5.1. Feature analysis

OpenCV's GoodFeaturesToTrack⁶² function was used to find the 10 strongest corner features in every single image for each image set. According to the function's documentation at the OpenCV's website:

“The function finds the corners with big eigenvalues in the image. The function first calculates the minimal eigenvalue for every source image pixel using the CornerMinEigenVal function and stores them in eigImage. Then it performs non-maxima suppression (only the local maxima in 3x3 neighborhood are retained). The next step rejects the corners with the minimal eigenvalue less than qualityLevel·max(eigImage(x,y)). Finally, the function ensures that the distance between any two corners is not smaller than minDistance. The weaker corners (with a smaller min eigenvalue) that are too close to the stronger corners are rejected.”

For each of the 10 strongest features, x and y coordinates are reported. For each image, the arithmetic mean of the 10 strongest features is calculated for x and y . Then an analysis of variance is performed to verify significant differences between the means of the three image sets.

The results of the statistical analysis show significant differences for the values of x among the three sets. However, for the values of y , there is significant difference between A and C, and between C and B, but the difference between A and B is not significant (see Table 25 and Figure 52).

⁶² http://docs.opencv.org/trunk/doc/py_tutorials/py_feature2d/py_shi_tomasi/py_shi_tomasi.html, visited on May 12th, 2012.

Comparisons for factor: **Sets within x**

Comparison	Diff of Means	t	P	P<0.05
C vs. B	31.022	38.204	<0.001	Yes
C vs. A	58.654	23.722	<0.001	Yes
B vs. A	27.633	10.932	<0.001	Yes

Comparisons for factor: **Sets within y**

Comparison	Diff of Means	t	P	P<0.05
C vs. A	165.090	66.768	<0.001	Yes
C vs. B	161.828	199.295	<0.001	Yes
B vs. A	3.263	1.291	0.197	No

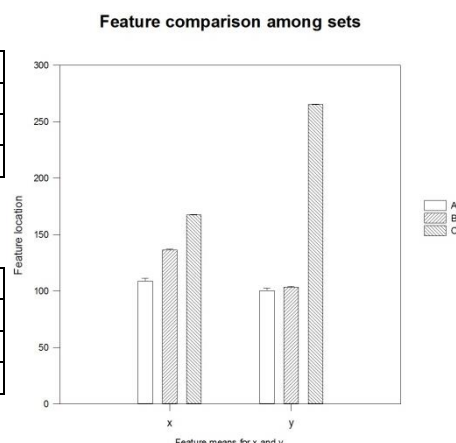


Table 25. ANOVA results for feature coordinates mean comparison.

Figure 52. Feature comparison among sets

These results somehow support the trends observed in the analysis of DORIS' performance in §5.5.5.2 above. A and B sets seem to share more features than the other two set combinations.

5.5.2.5.2. Description from the visual perception

From the mere visual perception, it can be said that there are two remarkable differences between set A and sets B and C: a) set A includes just photographs of technical artifacts, while sets B and C contain photographs but also diagrams or a combination of both; b) set A contains clean, noiseless images, while sets B and C often include the watermark artifact brand superimposed on the artifact photograph.

5.5.2.5.3. Definition of a prototypical image

According to the feature analysis and to the visual perception description above, an approximate definition of the prototypical image for an optimal performance of the BC model for comparable corpus location could be as follows:

The image that maximizes the BC model performance for artifact term translation is a photograph of a single artifact:

- 1) whose strongest feature's coordinate means do not have a statistical difference with the means calculated for the 10 strongest features of set B.
- 2) with a white background, where the object covers most of the photograph frame, with no surrounding or superimposed watermarks or text.

5.5.2.6. Evaluation through queries

Assuming that 1) images and terms in our database are correctly aligned, that is, that each image has associated the right term that describes the artifact in the image, and 2) that the images in our database feature the characteristics of our prototypical image, the results from the evaluation above in §5.5.2 should suffice to predict the performance of BC-Trans for artifact MWT translation. That is to say that having controlled these variables, a precision of 98.2% can be predicted as the performance of BC-Trans. In other words, the error of BC-Trans can be estimated as the sum of the error of the image matching task plus the error of the artifact term recognition for the image-term alignment task. On the other hand, if BC-Trans is fed with perfectly aligned image-term pairs in two languages, and images are prototypical, a precision of 98.2% in image-based term translation can be expected.

It was observed, however, that two of the three image sets analyzed in §5.5.2 do not entirely consist of prototypical images that maximize our model's performance and therefore do not yield as good results. While it is true that the characteristics of such non-prototypical image sets should not affect the validity of the BC model, it is also certain that a real world application of the model will probably face heterogeneous search spaces and queries that should also be considered.

This is why an additional evaluation of the system was carried out on the second use case presented above in §5.3.1, that is, query by image content. The system was manually queried with 200 labeled images of a different image set (set D) which visually seemed to comply with the requirements of the BC model prototypical image (see §5.2.5 above). This evaluation aims at measuring how many times a relevant image was retrieved from image set B when queried with images of set D. When a positive match occurred, the retrieved image also had an associated target term. This concretizes the term translation as well as the cross-language image indexing task.

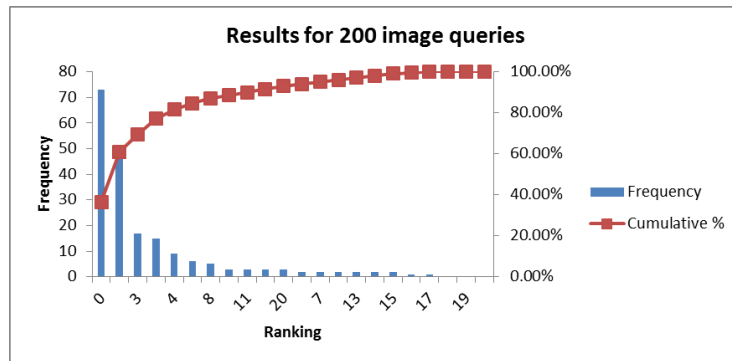


Figure 53. Evaluation through queries.

The evaluation results, summarized in Figure 53, suggest that the system should return to the user at least the 20 first matched images, and not just 3 as initially planned. If only the first three results are considered, the system would return relevant results for 41% of the queries. Considering the 20 first results, though, boosts overall precision up to 64%.

It was also observed that, as expected, DORIS' algorithm prioritizes texture over shape and color. That is the reason why some images yield better results than others. For instance, queries with images of a *rectifier* (see Figure 54), which is a highly-textured device, obtain relevant results most of the times. However, queries with images of a *bushing*, which generally presents an extremely even surface, rarely return relevant results.

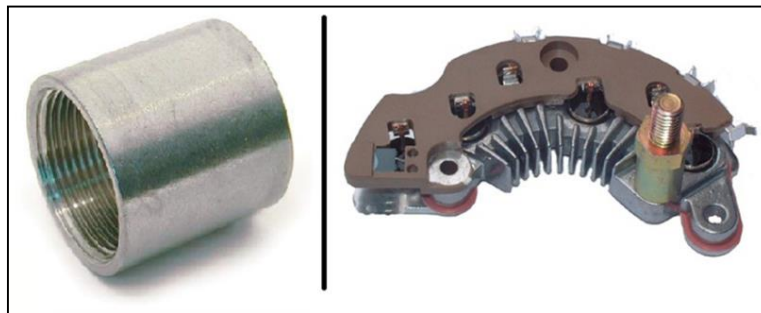


Figure 54. Two artifacts with different textures.

5.6. Remarks

The prototype has proved to be a promising application of the BC model. It uses a domain-oriented CBIR application (DORIS) to connect two related terms. Three different tasks were carried out with three different image sets to test DORIS' performance. The head noun matching (HNM) task applied to image set B yielded a MAP score of 0.655 and a 98.2% of

effectiveness. This percentage means that at least a 98.2% of the times the user obtains at least one relevant image among the three first results, which in turns means ones relevant target term, or, at least, one relevant target document.

Such an encouraging response from image set B suggests that images in this set probably have the right features to become the prototypical image for our model. The outstanding results of set B motivated a twofold analysis from both the visual perception and from a statistical analysis of variance to verify significant differences between the means of strong feature coordinates of the three image sets. The results of the analysis show a difference between set B and sets A and C. However, it seems that further analysis with other variables is necessary to determine which exactly the discriminant features that cause such good results are.

With the assumption that the visual component of the BC model contains the characteristics of a prototypical image and that the target search space contains relevant images with regard to the user's query, a high performance of the BC model for artifact term translation is expected. However, an additional evaluation of 200 queries with images of an independent image set shows that heterogeneity of image search spaces may lower precision to 64%.

6. CONCLUSIONS

Conclusions in this chapter have the form of a discussion that mixes summary of results, contributions, limitations, and future lines. We start by assessing the validation of our hypothesis and attainment of objectives as these constitute the spine of this work. Then, remarkable points are highlighted with regard to the BC model, the search space characterization, the image-term alignment main areas of work (i.e., multi-word term and artifact noun recognition), and the implemented prototype.

6.1. Hypothesis validation

In the Introduction chapter (§1.3.2), we introduced our hypothesis with the assumption that the bimodal co-occurrence of images and terms is natural to any discourse, in a greater or lesser extent, according to aspects which are inherent to each language's socio-cultural and/or technological resources. We also expected that such bimodal co-occurrence maximized its frequency in specialized technical documents. This first part of our hypothesis which corresponds to a monolingual setting was empirically proved by carrying out a study for English. This study was described in the Chapter 1 (§1.2.1), and also reported in (Burgos and Wanner, 2006).

Then, we moved on to a bilingual perspective of the hypothesis by predicting that at the very moment that the bimodal co-occurrence simultaneously exists in two documents of different languages (or even the same language) for an identical artifact referent, it can be stated that both images as well as both linguistic denominations designate the same artifact and, therefore, the terms are equivalent. This forecast in a bilingual setting was confirmed first by systematic manual observation and, then, by the application of the BC model in a functional prototype. Let us briefly examine three practical examples that illustratively summarize this first part of our hypothesis. The examples come from online catalogues of automotive spare parts in the search space defined in Chapter 3.

Example 1 (Table 26): This is an instance of a Spanish monolingual retrieval. The source image is designated by the index term *regulador Volvo 28v* while the target index term is *Regulador de voltaje*. If compared with the target term, the source term omits some modification (*de*

voltaje) and adds brand and tension specification. With regard to the images, both of them represent the same concept but with changes in their morphology and perspective.

Example 2 (Table 27): This is also an instance of a Spanish monolingual retrieval. The source image is designated by the index term *plaqueta rectificadora* while the target index term is *Portadiodo Mazda Hitachi Isuzu IHR-727*. If compared with the source term, the target term designates the concept with a synonym which features a completely different lexical and morphological configuration. The target term also adds brands and an additional reference code. With regard to the images, both of them represent the same concept but with changes in their morphology.

Example 3 (Table 28): This is an example of cross-language Spanish-English retrieval. The source image is designated by the index term *Portadiodo Mercedes Benz 366 Vr-904* while the target index term is *Rectifier*. If compared with the target term, the source term adds brands and an additional reference code. With regard to the images, both of them represent the same concept but with changes in their morphology.


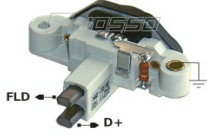
Source BC instance	Target BC instance																																																
<p>regulador Volvo 28v</p> 	<p>Información de Producto</p>  <p>RNB 311211 - Regulador de voltaje</p> <p>Especificaciones Técnicas:</p> <p>Tecnología: Electrónico Montaje: Incorporado Tensión: 12v Campo: 4A máx. Aro colector: 14mm Dist. orificios montaje: Circuito: A Nota 1: C/ resistor de 180 ohm Nota 2: Alternador 100A Nota 3:</p> <table border="1"> <thead> <tr> <th>Fabricante</th> <th>Nro. Original</th> <th>Aplicación de Vehículos</th> </tr> </thead> <tbody> <tr> <td>Bosch</td> <td>1197311211</td> <td>Mercedes-Benz -</td> </tr> <tr> <td>Bosch</td> <td>1197311213</td> <td>Volkswagen Passat</td> </tr> <tr> <td>Bosch</td> <td>1197311217</td> <td>Volvo -</td> </tr> <tr> <td>Bosch</td> <td>1197311219</td> <td></td> </tr> <tr> <td>Bosch</td> <td>1197311242</td> <td></td> </tr> <tr> <td>Citroën</td> <td>95644488</td> <td></td> </tr> <tr> <td>Citroën</td> <td>95644494</td> <td></td> </tr> <tr> <td>Fiat</td> <td>9944423</td> <td></td> </tr> <tr> <td>Fiat</td> <td>9950401</td> <td></td> </tr> <tr> <td>Lucas</td> <td>2131041</td> <td></td> </tr> <tr> <td>Lucas</td> <td>UCB419</td> <td></td> </tr> <tr> <td>Magneti Marelli</td> <td>940038016</td> <td></td> </tr> <tr> <td>Mercedes-Benz</td> <td>0021548106</td> <td></td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th>Parte Relacionada</th> <th>Fabricante</th> <th>Nro. Parte</th> </tr> </thead> <tbody> <tr> <td>Regulador de voltaje</td> <td>TCH</td> <td>05011</td> </tr> </tbody> </table>	Fabricante	Nro. Original	Aplicación de Vehículos	Bosch	1197311211	Mercedes-Benz -	Bosch	1197311213	Volkswagen Passat	Bosch	1197311217	Volvo -	Bosch	1197311219		Bosch	1197311242		Citroën	95644488		Citroën	95644494		Fiat	9944423		Fiat	9950401		Lucas	2131041		Lucas	UCB419		Magneti Marelli	940038016		Mercedes-Benz	0021548106		Parte Relacionada	Fabricante	Nro. Parte	Regulador de voltaje	TCH	05011
Fabricante	Nro. Original	Aplicación de Vehículos																																															
Bosch	1197311211	Mercedes-Benz -																																															
Bosch	1197311213	Volkswagen Passat																																															
Bosch	1197311217	Volvo -																																															
Bosch	1197311219																																																
Bosch	1197311242																																																
Citroën	95644488																																																
Citroën	95644494																																																
Fiat	9944423																																																
Fiat	9950401																																																
Lucas	2131041																																																
Lucas	UCB419																																																
Magneti Marelli	940038016																																																
Mercedes-Benz	0021548106																																																
Parte Relacionada	Fabricante	Nro. Parte																																															
Regulador de voltaje	TCH	05011																																															

Table 26. Spanish monolingual retrieval.


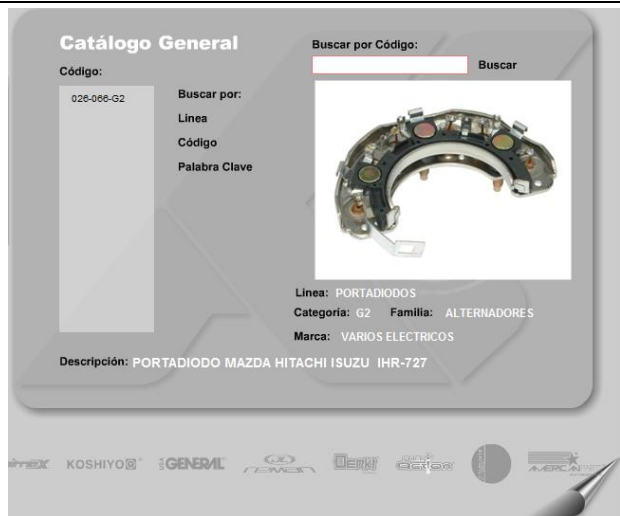
Source BC instance	Target BC instance
plaqueta rectificadora 	

Table 27. Spanish monolingual retrieval 2.


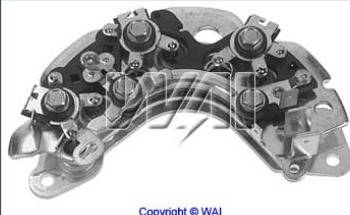
Source BC instance	Target BC instance																								
Portadiodo Mercedes Benz 366 vr-904 	<p>Home : Part Details</p>  <table border="1" data-bbox="614 1075 1276 1668"> <tr> <td>FR6023</td> <td>(31-212-1)</td> </tr> <tr> <td colspan="2">Rectifier</td> </tr> <tr> <td>For</td> <td>Ford 6G Series 130-135A* IR/IF Alternators</td> </tr> <tr> <td>Used On</td> <td>(2005-98) Ford, Lincoln, Mercury</td> </tr> <tr> <td>Replaces</td> <td>Ford F8AU-10A366-AAIP</td> </tr> <tr> <td>Unit Nos</td> <td>Ford 3W1U-10300-BB, 3W1Z-10346-BA; F8AU-10300-AB, -AC, -AD; F8AZ-10346-AB, F8ZU-10300-AC; XF2U-10300-BC, -BD; XF2Z-10346-BA</td> </tr> <tr> <td>Lester Nos</td> <td>7795, 8253, 8315</td> </tr> <tr> <td>Dimensions</td> <td>149mm Heat Sink OD</td> </tr> <tr> <td>Notes</td> <td>*FR6023 is the early design 8-diode rectifier for 6G stators with 2-leads/phase and Y-connection.</td> </tr> <tr> <td>Features</td> <td>Transpo 50A/300V press-fit diodes, OE validated Crimp-weld diode terminations Heavy duty copper connections reduce high-amp heat generation Pre-tinned terminals improves solder-ability Drop-down stator connection improves salvage rate for short-lead stators</td> </tr> <tr> <td>Also Consider</td> <td>FR6023SP Rectifier, Heavy Duty, Hi-Output Alts</td> </tr> <tr> <td>OE Bill Of Materials</td> <td>FR6023</td> </tr> </table>	FR6023	(31-212-1)	Rectifier		For	Ford 6G Series 130-135A* IR/IF Alternators	Used On	(2005-98) Ford, Lincoln, Mercury	Replaces	Ford F8AU-10A366-AAIP	Unit Nos	Ford 3W1U-10300-BB, 3W1Z-10346-BA; F8AU-10300-AB, -AC, -AD; F8AZ-10346-AB, F8ZU-10300-AC; XF2U-10300-BC, -BD; XF2Z-10346-BA	Lester Nos	7795, 8253, 8315	Dimensions	149mm Heat Sink OD	Notes	*FR6023 is the early design 8-diode rectifier for 6G stators with 2-leads/phase and Y-connection.	Features	Transpo 50A/300V press-fit diodes, OE validated Crimp-weld diode terminations Heavy duty copper connections reduce high-amp heat generation Pre-tinned terminals improves solder-ability Drop-down stator connection improves salvage rate for short-lead stators	Also Consider	FR6023SP Rectifier, Heavy Duty, Hi-Output Alts	OE Bill Of Materials	FR6023
FR6023	(31-212-1)																								
Rectifier																									
For	Ford 6G Series 130-135A* IR/IF Alternators																								
Used On	(2005-98) Ford, Lincoln, Mercury																								
Replaces	Ford F8AU-10A366-AAIP																								
Unit Nos	Ford 3W1U-10300-BB, 3W1Z-10346-BA; F8AU-10300-AB, -AC, -AD; F8AZ-10346-AB, F8ZU-10300-AC; XF2U-10300-BC, -BD; XF2Z-10346-BA																								
Lester Nos	7795, 8253, 8315																								
Dimensions	149mm Heat Sink OD																								
Notes	*FR6023 is the early design 8-diode rectifier for 6G stators with 2-leads/phase and Y-connection.																								
Features	Transpo 50A/300V press-fit diodes, OE validated Crimp-weld diode terminations Heavy duty copper connections reduce high-amp heat generation Pre-tinned terminals improves solder-ability Drop-down stator connection improves salvage rate for short-lead stators																								
Also Consider	FR6023SP Rectifier, Heavy Duty, Hi-Output Alts																								
OE Bill Of Materials	FR6023																								

Table 28. Cross-Language Spanish-English retrieval.

Last, in the third part of the hypothesis, we infer that equivalent multilingual terms would be located and retrieved by initially matching object representations (i.e., images) of artifacts, and that the nature of this interface would allow for the retrieval of equivalents with independence of their morphological, syntactic or lexical configuration. Indeed, it was possible to match equivalent terms via the artifact image as shown above, not only manually but also through the application of content-based image retrieval (CBIR) techniques. Moreover, we were able to use our image-based approach to find instances of terms whose equivalents were significantly different in their morphological, syntactic and lexical configuration not only between languages but also within the same language. We used one representative example of this case in the illustrations above where the Spanish term *portadiodo* has the English equivalent *rectifier*, but also has the Spanish synonym *placa rectificadora*.

We can conclude, then, that our hypothesis was entirely proved. It is true, as we will discuss below, that the efficient exploitation of this hypothesis depends to a great extent on the success and performance of a number of processes such as CBIR, image-term alignment, multi-word term (MWT) recognition, and noun classification. However, it is also certain that our hypothesis is independent by nature, and that the greater or lesser degree of achievement of the contributing areas should not undermine its power.

6.2. Objectives attainment

In this dissertation, we worked towards contributing to the compilation of wide-coverage dynamic multimodal terminological resources as raw material for translation- and terminology-based tasks. This objective was driven by the problem that a great deal of specialized translation and terminology-based tasks must be carried out on the basis of rather static low-coverage textual terminological resources, e.g., specialized dictionaries, terminological databases, etc.

The achievement of this general objective was pursued by addressing six specific goals introduced in Chapter 1:

- 1) *To propose a theoretical model and a practical implementation of a concept-based strategy to dynamically compile wide-coverage bimodal terminological dictionaries.* This thesis not only proposed a model for concept-based dictionary compilation but also contributed with a methodology and a functional prototype for its practical implementation. We showed how two documents in the same language or in different languages are theoretically interrelated by the bimodal co-occurrence (BC) hypothesis (see next objective) and how they can be connected in practice. The BC hypothesis gets the two

equivalent terms closer by making their corresponding images match. All together, the BC hypothesis, the theoretical background and properties of the linguistic and visual representations, as well as the interrelation among all these involved components make up what has been called here the bimodal co-occurrence (BC) model.

- 2) *To establish a bimodal co-occurrence model to align term and image.* This model was established through what was named here the bimodal co-occurrence (BC) hypothesis. The BC hypothesis assumes language independent bimodal co-occurrence of images and their designating term in the corpus. This implies that if the image of an object occurs in a document of the corpus, the corresponding term designating the object in the image will also occur in the same document. The BC hypothesis is also assumed to happen in a bilingual setting. That is, when there is an image of an artifact in the source language corpus along with its designating term, there should also be an image of the same artifact along with its designating term in the target language corpus. The practical alignment of images and terms in these settings is carried out using multi-word term (MWT) recognition and noun classification techniques that narrow down the set of designating candidates in collateral text.

- 3) *To analyze and characterize the search space which hypothetically will yield terminology to be used for dictionary compilation.* Chapter 3 shows how this objective was achieved. The characterization of our search space allowed for a clear delimitation of a web segment to be included in our study. After the analysis, it was also possible to determine the weight and representativeness in the corpus of the BC model components, that is, artifact images and MWTs. A qualitative and quantitative observation of these documents evidenced the potential as well as the drawbacks of the search space.

- 4) *To identify, propose and integrate techniques and tools which procedurally concretize the bimodal co-occurrence model.*⁶³ The characterization of the search space, the needs posed by our methodology, and the configuration of the BC model determined the techniques and tools that were necessary to be used in this research. With regard to the core techniques used here, we can group them in three categories:
 - a) *Low-level text processing.* For tokenization, lemmatization, and part-of-speech tagging the TreeTagger was used. Chunking was carried out with the TreeTagger and two python-based scripts.

⁶³ Tools and resources mentioned here are not cited when they have already been cited somewhere else in this dissertation.

- b) *High-level text processing*. For semantic annotation, the SuperSense Tagger (SST) and Freeling were used. For multi-word term extraction and noun classification, we applied the techniques described in Chapters 2 and 4.
- c) *Image matching*. For image matching, CBIR was used. CBIR was put into practice using DORIS (a Domain-ORiented Image Searcher implemented in Java).

Besides these core tools and techniques, other tools and resources were used for preprocessing tasks, observations, analysis, plotting, or for other supporting processes. For instance, two scripting languages were of extreme usefulness for text processing and web crawling: Perl and Python. Available packages and modules in both languages were used to implement *ad hoc* scripts. Python-based Natural Language Tool Kit (NLTK)⁶⁴ constituted a great springboard for a number of our scripts and for the implementation of our prototype. Likewise, Linux tools such as Sed, Grep, Awk, Geany, among others, were especially handy. For lexical semantics-related issues, WordNet and EuroWordNet were used and proved to be still valid and up to date. Our initial statistical analysis were carried out with SPSS® and Statgraphics®. Last, it is worth remarking the transverse contribution, utility, and power of regular expressions for low-level text processing tasks in this work.

- 5) *To evaluate the bimodal co-occurrence model for the specific task of building up bimodal terminological dictionaries*. Evaluation was presented in Chapter 4 for MWT recognition and artifact noun classification, and in Chapter 5 for image matching. For MWT recognition, we measured performance in various ways including analysis of variance and precision and recall. The results consistently show a better performance of the three experimented approaches when tested on English, although an acceptable performance was also reached for Spanish. The outcome of the analyses makes the predefined categories method the most effective one in the experiments.

As for artifact noun recognition, we examined three different approaches and evaluated them using discriminant analysis and precision and recall. The higher scores were reached with the lexical semantics-based approach for English (F=71) and with the Bayesian classifier for Spanish (F=74). It is expected, though, that the more specialized the text, the better the performance of the MWT recognition task. The domain plays an important role for the artifact MWT recognition task. The occurrence of terms and artifact nouns, as well as of related images, is a language- and domain-independent phenomenon, but it can be higher in some languages and in some domains. We expect that the application of these methods for artifact MWT recognition on appropriate

⁶⁴ <http://nltk.org/>

domains such as automotive online catalogs boosts the overall performance of a real world application.

For image matching, three different tasks were carried out in Chapter 5 with three different image sets to test CBIR performance. The head noun matching (HNM) task applied to an image set of optimal images yielded a MAP score of 0.655 and a 98.2% of effectiveness. However, an additional evaluation of 200 queries with images of a different image set shows that heterogeneity of image search spaces may lower precision to a 64%. Therefore, a high performance of the BC model for image-term alignment and artifact term translation is expected, provided that the visual component of the BC model contains the characteristics of our prototypical image and that the target search space contains relevant images with regard to the user's query,

- 6) *To design a functional prototype for the practical implementation of the BC model.* For a real world application of the BC model, a functional software prototype was implemented and described in Chapter 5. The software has been named BC-Trans (Bimodal Co-occurrence-Based Translation Software). It either finds translations for artifact multiword terms (MWTs) or finds terms for artifacts in photographic images. The user can type a term and get translations and images of his/her term or can upload a photographic image of an artifact and get the term for the artifact. Terms and images are stored in a server-based database. The user accesses BC-Trans and gets results via web browser. The system was designed on the basis of the BC Model, which means that it maximizes its performance when queried with artifact terms and/or photographic images.

6.3. Contributions

6.3.1. Characterization of the problem

The identification and systematic study of the problem that inspired this research can be considered the first contribution of this thesis. It has always been intuitively clear for translators, terminologists, and technical writers that current terminological repositories fall short of expectations for a number of tasks. However, a systematic analysis was necessary to quantify and characterize the claimed drawbacks of such resources. We provided such a study in Chapter 1 (§1.1) and with it the basis for our proposal was grounded. The study consisted of a corpus-driven dictionary analysis and a usage-based dictionary analysis. The former showed a discrepancy of term length between texts and dictionaries. According to the analysis, term length average in dictionaries is 2.11 for English and 2.86 for Spanish, which suggests a conservative approach to include mostly terms with a *noun noun* and *noun preposition noun* structure respectively. On the other hand, the usage-based dictionary analysis revealed that a

considerable number of dictionary entries retrieve very few documents from the Web. This finding illustrates well our problem since these little documented entries may have already become outdated or may be the product of a not very well informed translation.

6.3.2. The BC model

As a contribution to solve the whole problem, this thesis proposed and validated a language- and domain-independent model. The model shows how two documents that are theoretically interrelated by the BC hypothesis can be connected in practice. It takes advantage of the image-term bimodality that is recurrent in technical documents nowadays. While the BC model itself is already a contribution to solve the problem of image-term alignment and term translation, it is especially valuable for its application in specialized domains, since most of the existent proposals mainly deal with documents of general interest. Thus, even though there is research on medical image annotation and domain specific cross-language image retrieval (CLIR), as reported in Chapter 2, most of the work has been done in general domains. Likewise, the BC model enables user interaction, which widens its scope of applicability. That is, the BC model not only contributes to the research and professional activities of translators, terminologists, linguists, and semioticians, among others, but also gains importance in other settings such as e-commerce or online assistive technologies for it can support product search in digital bimodal databases.

It is also worth noting that the methodology established for the present research allows for the generation of comparable corpora as an important by-product. This can be achieved by two specific methodological moments, i.e., bilingual category matching in a web directory (see Chapter 3) and image matching on specific sites. Therefore, by matching two different-language categories in a web directory, a set of related documents in the involved languages can be linked, and by matching two images in two different documents, highly related contents are brought together. As an illustration, let us refer to the examples and figures in the hypothesis validation subsection above. It can be noticed that besides the target image and the target term in each case, there is also a good deal of additional related information contained in the target documents which can be compiled as a very helpful resource for a number of other tasks.

6.3.3. The Search Space

The World Wide Web is a huge and fast-growing space. The ease of access to the Web makes it possible for publishers and users to publish and read contents with few restrictions, but it also gives room for the dissemination of unreliable or poorly edited documents. Web directories were born as a criterion-based strategy to manually group and classify websites under categories and assure quality and reliability to a certain extent. It is the case of the Open Directory Project (ODP) which provided us with an initial departure point and delimitation for a suitable search space. However, beyond the broad categories and the structural description provided by the ODP, a finer characterization of the search space was necessary. This thesis offered a detailed insight into the map of a specific ODP category, namely, the *parts and accessories* category. For instance, the distribution of categories and subcategories in English and Spanish was determined. It was confirmed that English surpasses Spanish not only in number of websites but also, and more important, in the degree of granularity as for the definition of categories. Thus, for English, *automotive* is a non-leaf category, that is, under *automotive* the user can find some web sites but there are also further subcategories to keep browsing. On the contrary, for Spanish, *automotriz* is a leaf in most of the paths, i.e., it does not have subcategories but a set of sites classified under it.

The disparity of category distribution between languages such as the observed between English and Spanish has its effects. An analysis of frequencies showed that should there be the same category distribution between two languages, both sets of websites could be matched at the level of equivalent categories and therefore constitute an interesting source for bilingual comparable corpora. It is the case, for example, of English and Italian in the subdomain of automotive electrical parts, which can be reached by following common paths of the ODP (see Chapter 3, §3.2.2). A word frequency analysis of their keywords and descriptions under these categories confirms the close thematic relation between both corpora, which is not so fine-grained for the case of Spanish.

As another part of the search space characterization, an analysis of the URLs pointing to webpages under relevant categories of the ODP was performed. The strings in the URLs were analyzed to determine those URLs potentially containing instances of the bimodal co-occurrence (BC) hypothesis. From this analysis, we concluded that strings like *shop*, *products*, *store*, *catalogsearch*, *prod*, *catalog*, *product* and *part* for English, and strings like *catálogo*, *producto* and *shopping* for Spanish are frequent and seem to be characteristic of URLs containing instances

of the BC hypothesis. There are also some common productive strings for both languages like *product(s)* and *catalog*.

As a whole, the characterization of our search space informs on its relevance for the application of the BC model. A text and image representativeness analysis confirms the search space as a rich source of instances of the BC hypothesis, that is, image-term co-occurrences. It was also interesting to find out that even though the English number of tokens in the search space is considerably greater than Spanish (see §3.3.2.1.2), the difference in the number of types is not significant. The same relation was observed for nominal MWT candidates. The unbalance in the number of tokens and the similarity in the number of types suggest a larger industry sector in the English speaking countries on the one hand, and a well-established field (i.e., automotive engineering) in both languages, on the other hand.

It was also verified that, as part of the Web, the ODP is susceptible to the problems listed by Baeza-Yates and Ribeiro-Neto (1999, p. 368) concerning the data in the Web. Broadly speaking, those problems are related to data distribution, high percentage of volatile data, large volumes of data, unstructured and redundant data, quality of data, and heterogeneous data. Some specific instances of such problems are: a) misuse of capitalization, punctuation, and abbreviations, b) information redundancy, c) inaccessible data, d) heterogeneity of pages codes (Unicode, ANSI, etc.), e) diversity of languages, and f) use of natural language non-standard syntax.

6.3.4. MWT and artifact noun recognition

This thesis followed a rather naïve but practical text-based approach towards image-term alignment in a two-fold decision process, namely, anchor-based selection and search space reduction via MWT recognition and artifact noun classification. First, in Chapter 3 we found out that it is a common practice in catalogs to name the image file with the vendor's catalog reference code and to put such code also next to the artifact's denomination for the user's reference. We use this code as an anchor to align image and term in an efficient way. For more complex layouts where an anchor is not available, we narrow down the candidate terms in collateral text by means of MWT recognition and artifact noun classification. The result is a short list of candidate terms to be presented to the user as possible textual descriptors for the artifact in the image.

Rather than a genuine contribution to the tasks of MWT recognition and artifact noun classification, what we did here was to assess some of the approaches reported in the literature and their relevance according to the features of our search space, as summarized below.

6.3.4.1. MWT recognition

Two variables were decisive for this specific task: a) appropriateness of the corpus, and b) definition of term constituents. As for the appropriateness of the corpus, we depart from the assumption that compiling the appropriate data leverages MWT recognition from the very source, as proposed by Morin and Daille (2010). That is, the more specialized the corpus, the higher the relative frequency of MWTs, and the higher the probability of extracting relevant units. With this variable controlled, we were able to hypothesize that a high percentage of the extracted noun phrases (NPs) would be MWTs. This assumption motivated an experiment with a rule-based method using Quiroz's (2008) syntactic patterns (also used as baseline). However, even though Quiroz's patterns are terminological-like, precision was affected by the fact that they include an ample set of determiners and verbal forms. Thus, in order to reduce irrelevant candidates, this set was restricted. We gained precision with the expected decrease in recall, and the decision proved to be beneficial.

The decision process on what part-of-speech (POS) categories should a MWT contain raised interesting questions, though. For example, validated or referent MWTs (i.e., a gold standard) are used when evaluating precision and recall. These referent MWTs are defined manually or by means of a dictionary. Either way, the morphological configuration of the referent MWTs in dictionaries or gold standards sometimes poses more questions than answers and, of course, affects precision and recall. The big question here was again how a MWT is configured and what POS categories should constitute it. In some cases, it seems that the high frequency of use of certain sequences justifies their coexistence with pure terms in dictionaries, and that mere frequency may overcome the fact that two or more concepts lie in one single dictionary entry. In other words, our findings suggest that collocations are coexisting with terms in specialized dictionaries even when a single collocation may contain two independent terms.

As stated above, we found out that specialized dictionaries tend to be rather conservative as for the length of the MWTs they include (2.11 words for English and 2.86 words for Spanish). The difference in length between the two languages could be accounted for by the necessary usage of a preposition in Spanish for noun-noun modification, which would probably be “*de*”,

as shown in Chapter 4 (§4.2.1). On the other extreme, however, some instances of atypical terms as for their length and POS categories were found in dictionaries too. For example, the term *very small aperture terminal* found in the *Grand Dictionnaire Terminologique* features a gradable adverb and an evaluative adjective which are generally not expected to be part of a term.

Thus, the decisions made for MWT extraction aimed at reasonably relaxing the traditional conservative constraints of dictionaries, but also at controlling the extraction of atypical term candidates. We apply the same constraints to the other two tested methods: seeds and predefined categories.

The noisy nature of our search space as well as the non-standard use of language in many instances of relevant MWTs suggested the need of a syntax, typography, and morphology-independent method for MWT extraction. This is how we came up with the seed-based and the predefined categories (or bag of words) approaches. The seed approach, a lexical method, and the predefined categories approach, a part-of-speech-based method, have no statistical difference between them as performance concerns. The fact that these methods do not rely on a specific order of the constituents seems to overcome the problem of unexpected syntactic patterns. As expected, these two methods generate more noise than the rule-based approach, that is, less precision given by more irrelevant candidates.

It was interesting to see how the predefined categories (or bag of words) method adapted from Bourigault (1992) in conjunction with the selection of an appropriate corpus yielded the best results for both languages, even over a hard-to-beat baseline. Likewise, the few resources required by this method lowered the computational cost of processing with regard to the other two tested approaches.

6.3.4.2. Artifact noun recognition

The results obtained from our experiments seem to validate the trends reported in the literature. When local morpho-syntactic features are exhausted for concrete vs. abstract discrimination, a move towards semantic methods appears appropriate. This is when lexical semantics with support on resources such as WordNet and EuroWordNet as well as machine learning-based distributional semantics come into play as promising classification alternatives.

According to these trends, in this dissertation we conducted three different experiments for artifact noun classification: (1) a discriminant analysis using the number of images retrieved by concrete and abstract nouns, and the similarity between candidate terms and image names (for English, 74.4%); (2) a Bayesian model based on local linguistic features (for Spanish, 74%); and (3) semantic classification using lexical senses (for Spanish 70%, and for English 71-85.5%).

Although there is still much room for improvement of these results, it is also true that we are dealing with very noisy data which makes these scores promising. For example, the performance of the UKB and the SST could have been affected by the characteristics of the data set we used here, that is, a small test set made up of contexts from the telecommunications and assurance subject fields. While the telecommunications domain certainly has more artifact referents, the assurance domain uses more abstract referents which could affect the final recognition outcome. Evidence of the importance of the subject field in artifact noun classification can be found in an experiment we ran with 200 nouns taken from automotive engineering texts (§4.3.2.3) which yielded an overall accuracy of 85.5%.

6.3.5. Prototype

To make the most of the prevalence of today's multimodal digital documents, this thesis contributed not only with a theoretical model of the bimodal co-occurrence, but also with a web-based functional prototype that implements it. The software has been named *BC-Trans* (Bimodal Co-occurrence-Based Translation Software). It both finds translations for artifact terms and finds terms for artifacts in photographic images. The user either types a term to get translations and images of his/her term or uploads a photographic image of an artifact and gets the term for the artifact. Terms and images are stored in a server-based database. The user accesses BC-Trans and gets results via web browser. The system was designed on the basis of the BC Model, which means that it performs better when queried with artifact terms and/or photographic images.

The prototype uses a domain-oriented CBIR application (DORIS) to connect two related terms, and has proved to be a promising application of the BC model. Assuming a good performance in the image term alignment phase and that images follow the definition of our prototypical image, a precision of 98.2% can be expected. With more heterogeneous image sets and with higher variability in image features, though, performance may decrease down to 64%.

A definition of a prototypical image was also provided. It was proved that images used in online catalogs follow low level and visual patterns that optimize the prototype's performance. For this definition, the means of the 10 strongest features of three different image sets were statistically calculated and analyzed. In addition to this, the prototypical image was visually described as one with a white background, where the object covers most of the photograph frame, with no surrounding or superimposed watermarks or text.

6.4. Limitations

6.4.1. The BC model

The present thesis focuses on the bimodal nature of technical-scientific documents where images and their relation with terms are of great importance for specialized communication. While this bimodal co-occurrence means the greatest strength of our proposal, it also implies its main limitation for the BC model is constrained to deal only with artifact nouns and maximizes its performance in certain domains. The problem described in Chapter 1, of course, not only deals with concrete, artifact nouns, but also with communication, event, process, and other types of abstract nouns and even with other parts of speech, not only nouns. Likewise specialized communication takes place also in other knowledge areas including those unsuitable for the application of our model.

It is also true that, being the present proposal an interdisciplinary one, the BC model undergoes the limitations contributed by each of the involved disciplines, i.e., CBIR, MWT recognition, and noun classification. For example, our CBIR application has proved to perform very well when dealing with images of the characteristics defined in Chapter 5 (§5.5.2.5.3) reaching a precision of 98.2. However, performance decreases with non-prototypical images. Therefore, a more robust behavior by the system depends on the progress of CBIR and term recognition techniques.

It is encouraging, however, that the BC model as a whole will be automatically enhanced with the independent developments of the involved disciplines as new advances can be incorporated to the model. For instance, it is expected that CBIR keeps evolving and efficiently narrowing down the divide between low-level features and the semantic interpretation of images. This is especially relevant for the BC model if descriptors and techniques are proposed for images with the characteristics of the prototypical image or other images in relevant search spaces.

6.4.2. MWT and artifact noun recognition

Likewise, there are limitations for the MWT recognition task which can be qualified as external and internal limitations. By external limitations we mean that current techniques for MWT extraction are far from being perfect. A *non plus ultra* point seems to have been reached in the search for innovative paradigms for term extraction. The reported strategies recurrently revolve around the existent paradigms, i.e., linguistic, statistical, and hybrid approaches. New proposals consist more of creative combinations of statistical techniques and linguistic features. The possibilities for combination are numerous and interesting, though, and we are then constrained by such paradigms and techniques. Secondly, we are internally limited on the one hand by the fact that we did not deeply experiment here with statistical or hybrid approaches for MWT recognition, and on the other hand by the characteristics of our data. It has been proved by previous research that hybrid methods may boost term recognition. Nonetheless, in order to take advantage of them for our purposes, it is necessary to find first the right strategies to statistically model the noisy configuration and distribution of terms in our search space.

6.4.3. Prototype

In the same line, being the proposed software prototype described in Chapter 5 a practical implementation of the BC model, it inherits the same limitations of the model and therefore intends to be a complement or to be complemented by other resources that share common goals. As a software product, it presents also other technical limitations. For instance, its web interface enables the user to query the pre-loaded databases but does not allow for building and indexing a new database. It is a reasonable limitation since BC-Trans is a web-based service which makes the upload of thousands of images for database construction not possible nowadays in a reasonable period of time. Similarly, the software may still be improved by implementing more sophisticated search methods, like text-based fuzzy searches or mixed (visual and text) methods for relevant image-term pair retrieval.

6.5. Lines of future research

The limitations presented above derive either from time and resource constrains or from being out the immediate scope of the present research. They therefore constitute interesting matter for future research given their potential to make the proposal of this thesis more robust and its application more generalizable. For instance, with regard to MWT recognition, although we proved that a combination of a suitable corpus and a linguistic technique yields encouraging

results, an interesting future line should include more sophisticated statistical and hybrid approaches. The knowledge derived from the search space characterization as well as from MWT description and artifact noun patterns constitute a very informative input for statistical and machine learning methods for both MWT and artifact noun recognition. A combination of this knowledge with other successful strategies reported in the literature (e.g., Peirsman et al., 2008 and Van de Cruys, 2008) could help improve results while dealing with a very noisy and problematic corpus.

There are also interesting perspectives related to assessment techniques. Our approaches here for MWT recognition and artifact noun classification are tested and evaluated independently. An interaction of the three methods for each task is still pending in order to see how they can contribute to each other in a voting-like strategy for better performance. In addition to this, more corpus-driven and dictionary-based description of terms as well as clear criteria for term recognition are also necessary in order to define suitable gold standards for term recognition evaluation. Such criteria should help demarcating the boundaries between terms and other non-terminological expressions included in dictionaries (e.g., Microsoft® glossaries for software localization⁶⁵, where terms coexist with complete sentences as entries).

With the accomplishment of the research lines previously proposed, the software prototype will also automatically benefit. There are, however, other thinkable improvements to enhance performance and user experience such as a) creating a downloadable application so the user can locally generate the database file with his/her own photographs and then upload it for further use, b) combining text and visual methods for image retrieval, c) adding new languages and domains, d) enabling fuzzy search for terms, e) including (pseudo)relevance feedback for terms and images, and f) activating term record download.

Last, as the BC model is addressed to artifact nouns and mostly applicable to certain subject fields, it is susceptible of being extended to cover other noun categories, parts of speech and domains. Integrating other components into the BC model or the BC model into other systems would set up a more comprehensive environment for translation, terminography, technical writing, and other related areas.

⁶⁵ <https://www.microsoft.com/Language/en-US/Terminology.aspx>

REFERENCES

- Agirre, E. & Soroa, A. (2009). Personalizing Pagerank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Athens, Greece.
- Ahmad, K.; Vrusias, B. & Tariq, M. (2002). Co-Operative Neural Networks and Integrated Classification. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN'02)*, IEEE Press, 1546-1551.
- Ai'khenval'd, A. (2000). *Classifiers: A typology of noun categorization devices*. England: Oxford University Press.
- Altarriba, J.; Bauer, L. M. & Benvenuto, C. (1999). Concreteness, Context Availability, and Imageability Ratings and Word Associations for Abstract, Concrete, and Emotion Words. *Behavior Research Methods, Instruments, & Computers* 31(4), 578-602.
- Alvarez, C.; Oumohmed, A. I.; Mignotte, M. & Nie, J.-Y. (2005). Multilingual Information Access for Text, Speech and Images. *Toward Cross-Language and Cross-Media Image Retrieval*. Heidelberg, Berlin: Springer Berlin. 676-687.
- Armour, F. & Miller, G. (2001). *Advanced Use Case Modeling: Software Systems*. EE .UU: Addison-Wesley.
- Axenopoulos, A.; Daras, P.; Malassiotis, S.; Croce, V.; Lazzaro, M.; Etzold, J.; Grimm, P.; Massari, A.; Camurri, A.; Steiner, T. & Tzovaras, D. (2012). I-SEARCH: A Unified Framework for Multimodal Search and Retrieval. *The Future Internet*. Heidelberg: Springer Berlin, 130-141.
- Baeza-Yates, R. & Ribeiro-Neto. (1999). *Modern Information Retrieval*. Addison-Wesley: Longman Publishing co.
- Baldwin, T.; Kim, S. N.; Indurkha, N. & Damerau, F. J. (2010). *Multinword Expressions*. CRC Press, Taylor and Francis Group: Boca Raton, FL.
- Barbu, E. (2008). Combining methods to learn feature-norm-like concept descriptions. *Proceedings of Workshop Lexical semantics: bridging the gap between semantic theory and computational simulations Hamburg*, August 4-15, 2008.
- Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, 1313-1316.
- Bay, H.; Ess, A.; Tuytelaars, T. & Gool, L. (2008). Speeded-up robust feature. *Computer Vision and Image Understanding*. 110(3), 346-359.
- Bel, N.; Espeja, S. & Marimon, M. (2008). Automatic acquisition for low frequency lexical items. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

- Besançon, R. & Millet, C. (2006). Using Text and Image Retrieval Systems: Lic2m Experiments at ImageCLEF 2006. In *Working notes of the CLEF 2006 Workshop*.
- Besançon, R.; Ferret, O.; Fluhr, C.; Peters, C.; Clough, P.; Gonzalo, J.; Jones, G. J. F.; Kluck, M. & Magnini, B. (2005). Multilingual Information Access for Text, Speech and Images. *Integrating New Languages in a Multilingual Search System Based on a Deep Linguistic Analysis*. Springer Berlin / Heidelberg. 83-89.
- Biber, D.; Johansson, S.; Leech, G.; Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman: EE. UU.
- Bloom, P. (2000). *How Children Learn the Meanings of Words*. MIT Press: EE. UU.
- Bonin, F.; Dell'Orletta, F.; Venturi, G. & Montemagni, S. (2010), Contrastive Filtering of Domain-Specific Multi-Word Terms from Different Types of Corpora. In *Proceedings of the Multword Expressions: From Theory to Applications (MWE 2010)*, Beijing, pp. 77–80.
- Bosque, I. (1999). Gramática descriptiva de la lengua castellana. *El nombre común*. Espasa: Calpe. pp. 3-75.
- Bourigault, D. (1992). Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *COLING*, 977-981.
- Bullinaria, J. A. (2008). Semantic Categorization Using Simple Word Co-occurrence Statistics. In Baroni Marco; Evert Stefan & Lenci Alessandro, eds. *ESSLLI Workshop on Distributional Lexical Semantics*.
- Burgos, D. & Wanner, L. (2006). Using CBIR for Multilingual Terminology Glossary Compilation and Cross-Language Image Indexing. In *Proceedings of the Workshop on Language Resources for Content-based Image Retrieval*. 5-8.
- Burgos, D. (2009). Clasificación de nombres concretos y abstractos para extracción terminológica. In *La terminología y los usuarios de la información: puntos de encuentro y relaciones necesarias para la transferencia de la información*.
- Cabré, M. (1993). *La terminología: teoría, metodología, aplicaciones*. Antártida: Argentina.
- Cabré, M. (2000). Elements for a theory of terminology: Towards an alternative paradigm, *Terminology*. 6(1), 35-57.
- Cabré, M. (2003). Theories of terminology. Their description, prescription and explanation, *Terminology*. 9(2), 163-199.
- Carreras, X.; Chao, I.; Padró, L. & Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*.
- Carson, C.; Belongie, S.; Greenspan, H. & Malik, J. (2002). Blobworld: Image Segmentation Using Expectation-Maximisation and its Application to Image Querying. *IEEE Trans. Pattern Analysis and Machine Intelligence*. 24(8), 1026-1038.

- Chang, S.; Smith, J.; Beigi, M. & Benitez, A. (1997). Visual Information Retrieval from Large Distributed Online Repositories. *Communications of the ACM* 40(12), 63-71.
- Chang, Y.-C. & Chen, H.-H. (2006). Approaches of Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval. In *Working notes of the CLEF 2006 Workshop*.
- Ciaramita, M. & Altun, Y. (2006). Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Clough, P. (2005). Multilingual Information Access for Text, Speech and Images, Springer Berlin / Heidelberg, Berlin, *Caption and Query Translation for Cross-Language Image Retrieval*, 614-625.
- Clough, P.; Grubinger, M.; Deselaers, T.; Hanbury, A. & Müller, H. (2006). Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In *Working notes of the CLEF 2006 Workshop*.
- Craig, C. (1986). *Noun classes and categorization*. John Benjamins: Philadelphia.
- Datta, R.; Joshi, D.; Li, J. & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* 40(2), 1-60.
- Daumke, P.; Paetzold, J. & Markó, K. (2006). Morphosaurus in ImageCLEF 2006: The effect of subwords on biomedical IR. In *Working notes of the CLEF 2006 Workshop*.
- De Coster, J. (2003). *Dictionary for Automotive Engineering*. München: K.G. Saur.
- Donnell, J. (2005). Illustration and Language in Technical Communication. *Journal of Technical Writing and Communication* 35(3), 239-271.
- Drouin, P. (2003). Term Extraction Using non-Technical Corpora as a Point of Leverage, *Terminology* 9(1), 99-115.
- Eco, U. (1989). *The Open Work*. Cambridge, Mass.: Harvard University Press.
- Esselink, B. (2000). *A Practical Guide to Localization*. John Benjamins: Philadelphia.
- Estopà, R. (1999). *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. PhD thesis, Barcelona: Institut Universitari de Lingüística Aplicada.
- Estopà, R.; Cabré, M.; Bach, C. & Martí, J. (2007). *Terminología y derecho: complejidad de la comunicación multilingüe*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra i Documenta Universitària, El problema de la identificación y la delimitación de unidades terminológicas en contexto.
- Estopà, R.; Vivaldi, J. & Cabré, M. T. (2000). Use of Greek and Latin Forms for Term Detection. In *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC-2000)*. Athens, 855-859.

- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press: Cambridge.
- Feng, Y. & Lapata, M. (2008). Automatic Image Annotation Using Auxiliary Text Information. In *ACL HLT*.
- Fillmore, C. J. (1977). Linguistic Structures Processing, Amsterdam. *Scenes-and-Frames Semantics*. N. Holland. 55-88.
- Freixa, J. (2002). *La variació terminològica: anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient*. PhD thesis. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
- Gamero, S. (2001). *La traducció de textos tècnics*. Barcelona: Ariel.
- Gelasca, E. D.; Ghosh, P.; Moxley, E.; Guzman, J. D.; Xu, J.; Bi, Z.; Gauglitz, S.; Rahimi, A. M. & Manjunath, B. S. (2007). *CORTINA: Searching a 10 Million + Images Database*.
- Geradts, Z. (2003). *Content-Based Information Retrieval from Forensic Image Databases*. PhD thesis. University of Utrecht, Netherlands.
- Grubinger, M.; Clough, P.; Hanbury, A. & Müller, H. (2007). Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task. In *Working Notes for the CLEF 2011 Workshop*.
- Grubinger, M.; Clough, P.; Müller, H. & Deselears, T. (2006). The IAPR-TC12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage 2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06*, 13-23.
- Hackl, R. & Mandl, T. (2005). Mono- and Bilingual Retrieval Experiments with a Social Science Document Corpus. In *WORKING NOTES CLEF 2005 Workshop*.
- IEEE (2009). *Standard for Information Technology—Systems Design— Software Design Descriptions(1016)*. IEEE, Technical report, IEEE.
- International Electrotechnical Commission. (2005). *International Electrotechnical Vocabulary (IEV)* online database. Geneva. <http://domino.iec.ch/iev/iev.nsf/Welcome?OpenForm>
- Iqbal, I. & Aggarwal, J. K. (2003). Feature Integration, Multi-image Queries and Relevance Feedback in Image Retrieval. In *6th International Conference on Visual Information Systems (VISUAL 2003)*, 467-474.
- ISO TC 37/SC4 N021. (2002). *ISO TC 37/SC4 N021*. Basic Requirements for Terminology Management within ISO Technical Committees'.
- Jacobson, I. (1992). *Object-Oriented Software Engineering: A Use Case Driven Approach*. ACM Press (AddisonWesley Pub): New York.
- Jacquemin, C.; Klavans, J. L. & Tzoukermann, E. (1997). Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. In *Proceedings of the 35th Annual Meeting of the ACL*.

- Jaramillo, G. & Branch, J. (2009). Recuperación de Imágenes por Contenido Utilizando Momentos. *Revista Iteckne* 5(2). 100-103.
- Jaramillo, G. E. & Branch, J. W. (2009). Recuperación Eficiente de Información Visual Utilizando Momentos. In *XXXV Conferencia Latinoamericana de Informática - CLEI 2009*.
- Judea, A. (2013), 'Unsupervised Training Set Generation for Automatic Terminology Acquisition', Technical report, Institute for Natural Language Processing, University of Stuttgart.
- Justeson, J. & Katz, S. (1995). Technical Terminology: some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering* 1, 9-27.
- Kageura, K. & Umino, B. (1996). Methods of automatic term recognition: A review, *Terminology*. 9(2), 259-289.
- Katrenko, S. & Adriaans, P. (2008). Qualia Structures and their Impact on the Concrete Noun Categorization Task. In *Proceedings of the "Bridging the gap between semantic theory and computational simulations" workshop at ESSLLI 2008*.
- Kußmaul, P. (2005). Translation through Visualization. *Meta* 50(2), 378-391.
- Lacoste, C.; Chevallet, J.-P.; Lim, J.-H.; Wei, X.; Raccoceanu, D.; Hoang, D. L. T.; Teodorescu, R. & Vuillenemot, N. (2006). IPAL Knowledge-based Medical Image Retrieval in ImageCLEFmed 2006. In *Working notes of the CLEF 2006 Workshop*.
- Lam-Adesina, A. M. & F.Jones, G. J.Peters, C.; Gonzalo, J.; Braschler, M. & Kluck, M. (2004). Comparative Evaluation of Multilingual Information Access Systems, Springer Berlin / Heidelberg. *Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval*, 271-285.
- Lam-Adesina, A. M. & Jones, G. J. F. (2003). Advances in Cross-Language Information Retrieval, Springer Berlin / Heidelberg, *EXETER AT CLEF 2002: Experiments with Machine Translation for Monolingual and Bilingual Retrieval*. 127-146.
- Larson, R. R. (2006). Domain Specific Retrieval: Back to Basics. In Peters C. Nardi, A. & J.L. Vicedo, eds. *WORKING NOTES CLEF 2006 Workshop, 20-22 September, Alicante, Spain*.
- Larson, R. R. (2006). Text Retrieval and Blind Feedback for the ImageCLEF Photo Task. In *Working notes of the CLEF 2006 Workshop*.
- Lavrenko, R.; Manmatha & Jeon, J. (2003). A model for learning the semantics of pictures. In *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems, Vancouver, BC*.
- Leong, B. (2012). *Modeling Synergistic Relationships between Words and Images*. PhD Thesis. University of North Texas.

- Li, W. & McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the International Conference on Machine learning*.
- Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR'99)*.
- Manjunath, B. & Salembier, P. (2002). *Introduction to MPEG-7: Multimedia Content Description Language*. Wiley: EE. UU.
- Maynard, D. & Ananiadou, S. (1999). Identifying Contextual Information for Multi-Word Term Extraction. In: Sandrini, Peter (ed.) *Terminology and Knowledge Engineering (TKE '99)*. Innsbruck. Wien: TermNet. 212-221.
- McCarthy, D.; Koeling, R. & Carroll, J. (2004). Finding Predominant Senses in Untagged Text. In *Proceedings of ACL 2004*.
- Mikolajczyk, K. & Schmid, C. (2005). Performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 27(10), 1615–1630.
- Miller, G.C. (1998). WordNet: An Electronic Lexical Database (Language, Speech, and Communication). *Nouns in WordNet*. The MIT Press: Cambridge.
- Monterde, A. M. (2002). *Interrelaciones e interdependencias entre distintas formas de representación conceptual: Estudio en tres niveles de especialización en textos sobre instalaciones de combustible de aviones*. PhD thesis, University of Las Palmas de Gran Canaria.
- Monterde, A. M.; Gonzalo G. & García, V. (2004). Manual de documentación y terminología para la traducción especializada. *Importancia de la ilustración para la traducción técnica: estudio en el campo de la aeronáutica*. Madrid: Arco / Libros. 259-274.
- Morin, E. & Kageura, K. (2010). Brains, not Brawn: The Use of “smart” Comparable Corpora in Bilingual Terminology Mining. *ACM Trans. Speech Lang, Process.* 7. Article 1.
- Müller, H.; Kalpathy-Cramer, J.; Hatt, W.; Bedrick, S. & Hersh, W. (2008). Overview of the ImageCLEFmed 2008 Medical Image Retrieval Task. In *Working Notes for the CLEF 2008 Workshop*.
- Nazar, R. (2011). A Statistical Approach to Term Extraction. *International Journal of English Studies*. 11(2), 159-182.
- Nazar, R.; Vivaldi, J. & Wanner, L. (2012). Automatic Taxonomy Extraction for Specialized Domains Using Distributional Semantics. *Terminology*. 18(2), 188-225.
- Ortega, M. L. (2002). Una propuesta para el análisis de las imágenes científicas en la formación del profesorado: una aproximación socio-epistemológica. *Investigación y Desarrollo*. 10(001), 76-99.
- Orueta, G. (2004). *Diccionario del motor*. Madrid: CIE / Dossat.
- Osinski, D. (2004). Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data, in *Advances in Soft Computing, Intelligent Information*

Processing and Web Mining. *Proceedings of the International IIS: IIPWM '04 Conference, Zakopane, Poland*, 369-378.

- Oxford Dictionaries. (2010). Oxford Dictionaries. Oxford University Press: Oxford. <http://oxforddictionaries.com/definition/coil> (accessed: January 29, 2012).
- Paek, S.; Sable, C. L.; Hatzivassiloglou, V.; Jaimes, A.; H., S. B.; Chang, S.-F. & McKeown, K. (1999). Integration of Visual and Text Based Approaches for the Content Labeling and Classification of Photographs. In *ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval*.
- Pentland, A.; Picard, R. & Sclaro, S. (1996). Photobook: Content-Based Manipulation of Image Databases. *International Journal of Computer Vision*. 18(2), 233-254.
- Perugini, S. (2008). Symbolic links in the Open Directory Project. *Information Processing and Management*. 1-21.
- Petras, V. Peters, C.; Clough, P.; Gonzalo, J.; Jones, G. J. F.; Kluck, M. & Magnini, B. (2005). Multilingual Information Access for Text, Speech and Images. *GIRT and the Use of Subject Metadata for Retrieval* Springer Berlin / Heidelberg. 298-309.
- Popescu, A.; Tsirikia, T. & Kludas, J. (2010). Overview of the Wikipedia Retrieval Task at ImageCLEF 2010. In *Working Notes for the CLEF 2010 Workshop*.
- Pustejovsky, J. & Meaning, P. (2001). The Syntax of Word. *Type Construction and the Logic of Concepts*. Cambridge University Press: Cambridge.
- Quiroz, G. (2008). *Los sintagmas nominales extensos especializados en inglés y en español: Descripción y clasificación en un corpus de genoma*. PhD thesis, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.
- Real Academia Española. (2010). *Nueva gramática de la lengua española*. Madrid: Espasa Libros.
- (1998). *Routledge English Technical Dictionary*. London: Routledge.
- (1998). *Routledge Spanish Dictionary of Business, Commerce, and Finance*. New York: Routledge.
- Rui, Y.; Huang, T. S. & Chang, S. (1999). Image Retrieval: Current Techniques, Promising Directions and Open Issues. *Journal of Visual Communication and Image Representation*. 10, 39-62.
- Santamaría, C.; Verdejo, J. G. & Verdejo, F. (2003). Automatic Association of Web Directories to Word Senses. *Computational Linguistics*. 29(3), 485-502.
- Savary, A.; Jacquemin, C. & Grefenstette, G. (2003). Text- and Speech-Triggered Information Access, Springer. *Reducing Information Variation in Text*. 145-181.
- Schmid, H. (1994). Probabilistic part-of-speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*. 44-49.

- Smith, J. R. & Chang, S.-F. (1996). VisualSEEK: a Fully Automated Content-Based Image Query System. In *Proceedings ACM International Conference Multimedia*. 87-98.
- South, D. W. & Dwiggins, B. H. (1999). *Diccionario de automoción*. Madrid: Paraninfo.
- Srihari, R. K. & Burhans, D. (1994). Visual Semantics: Extracting Visual Information from Text Accompanying Pictures. In *AAAI 94*.
- Teague, M. (1980). Image Analysis via the General Theory of Moments. *Journal Opt. Society American*. 70(8), 920-930.
- Tercedor-Sánchez, M. I. & Abadía-Molina, F. (2005). The Role of Images in the Translation of Technical and Scientific Texts. *Meta* 50, 1-7.
- The Open Directory Project. (2002). *The Open Directory Project*.
- Tollmar, K.; Yeh, T. & Darrell, T. (2004). IDEIXIS: Image-based Deixis for Finding Location-Based Information. In *CHI '04 extended abstracts on Human factors in computing systems*. 781-782.
- Tsikrika, T.; Popescu, A. & Kludas, J. (2011). Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011. In *Working Notes for the CLEF 2011 Workshop*.
- Urcid, P. (2003). *Búsqueda de imágenes por contenido en bibliotecas digitales*. Thesis. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas, Puebla.
- Villegas, M. & Paredes, R. (2012). Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task. In *CLEF 2012 working notes*.
- Vivaldi, J. & Rodríguez, H. (2002). Medical Term Extraction using the EWN Ontology, in TKE 2002. *Terminology and Knowledge Engineering Proceedings. 6th International Conference 28th-30th*. 137-142.
- Vivaldi, J. & Rodríguez, H. (2007). Evaluation of Terms and Term Extraction Systems: A Practical Approach. *Terminology*. 13(2), 225-248.
- Vivaldi, J. (2001). *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. PhD thesis, Barcelona: Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya.
- Vivaldi, J. (2003). *Sistema de reconocimiento de términos Mercedes. Manual de utilización*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
- Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.
- Vrochidis, A.; Moumtzidou, I. & Kompatsiaris (2012). Concept-based Patent Image Retrieval. *World Patent Information Journal* 34(4), 292-303.

- Vrochidis, A.; Moutzidou, S. and Kompatsiaris, I. (2012). Concept-based Patent Image Retrieval. *World Patent Information Journal*. 34(4), 292-303.
- Winder, S. & Brown, M. (2008). Learning Local Image Descriptors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*.
- Winder, S.; Hua, G. & Brown, M. (2009). Picking the Best Daisy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*.
- Yi, W.; Zhuang, Y. & Pan, Y. H. (2000). Image Retrieval System for Web: Webscope-CBIR. In *Proceedings of the 11th International Workshop on Database and Expert Systems Applications*. 620-624.
- Yves, K. & Geeraerts, D. (2008). Size Matters: Tight and Loose Context Definitions in English Word Space Models. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*.
- Zhang, Z.; Iria, J.; Brewster, C. & Ciravegna, F. (2008), A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*, Marrakech, Morocco.
- Ziering, P.; van der Plas, L. & Schütze, H. (2013), Multilingual Lexicon Bootstrapping - Improving a Lexicon Induction System Using a Parallel Corpus. In *Proceedings of the International Joint Conference on Natural Language Processing*, Nagoya, Japan, pp. 844–848.
- Ziering, P.; van der Plas, L. & Schütze, H. (2013), Bootstrapping Semantic Lexicons for Technical Domains. In *Proceedings of the International Joint Conference on Natural Language Processing*, Nagoya, Japan, pp. 1321–1329.

ANNEXES

Annex 1. English and Spanish syntactic patterns

English:⁶⁶

1. JJ + JJ + JJ + NN + NN
2. JJ + JJ + JJ + NN
3. JJ + JJ + NN + NN + NN
4. JJ + JJ + NN + NN
5. JJ + JJ + NN
6. JJ + NN + JJ + NN
7. JJ + NN + NN + NN + NN
8. JJ + NN + NN + NN
9. JJ + NN + NN
10. JJ + NN
11. NN + JJ + JJ + NN
12. NN + JJ + NN + NN
13. NN + JJ + NN
14. NN + NN + JJ + NN
15. NN + NN + NN + NN + NN
16. NN + NN + NN + NN
17. NN + NN + NN
18. NN + NN
19. RB + JJ + JJ + NN
20. RB + JJ + NN + NN + NN
21. RB + JJ + NN + NN
22. RB + JJ + NN

Spanish:

1. NC + ADJ
2. NC + ADJ + ADJ
3. NC + ADJ + ADJ + ADJ
4. NC + ADJ + ADJ + NC
5. NC + ADJ + ADJ + PREP + NC
6. NC + ADJ + ADV + ADJ
7. NC + ADJ + ADV + VLadj
8. NC + ADJ + ADV + VLadj + PREP + NC
9. NC + ADJ + NC
10. NC + ADJ + PREP + ADJ
11. NC + ADJ + PREP + ADJ + NC
12. NC + ADJ + PREP + ADJ + NC + ADJ
13. NC + ADJ + PREP + ADJ + NC + PREP + NC
14. NC + ADJ + PREP + NC
15. NC + ADJ + PREP + NC + ADJ
16. NC + ADJ + PREP + NC + ADV + ADJ
17. NC + ADJ + PREP + NC + NC
18. NC + ADJ + PREP + NC + PREP + NC
19. NC + ADJ + PREP + NC + PREP + NC + ADJ
20. NC + ADJ + PREP + NC + PREP + NC + PREP + NC
21. NC + ADJ + PREP + NC + VLadj
22. NC + ADJ + VLadj
23. NC + ADJ + VLadj + PREP + NC
24. NC + ADJ + VLadj + PREP + NC + ADJ
25. NC + ADV + ADJ

⁶⁶ For a description of the used tagset, go to <http://www.ims.uni-stuttgart.de/projekte/complex/TreeTagger/>.

26. NC + NC
27. NC + NC + ADJ
28. NC + NC + NC
29. NC + NC + PREP + NC
30. NC + NC + PREP + NC + ADJ
31. NC + NC + PREP + NC + PREP + NC
32. NC + NC + PREP + NC + PREP + NC + ADJ
33. NC + PREP + ADJ
34. NC + PREP + ADJ + ADJ
35. NC + PREP + ADJ + ADJ + ADJ
36. NC + PREP + ADJ + ADJ + PREP + NC
37. NC + PREP + ADJ + ADJ + PREP + NC + ADJ
38. NC + PREP + ADJ + CC
39. NC + PREP + ADJ + NC
40. NC + PREP + ADJ + NC + ADJ
41. NC + PREP + ADJ + NC + PREP + NC
42. NC + PREP + ADJ + NC + PREP + NC + ADJ
43. NC + PREP + NC
44. NC + PREP + NC + ADJ
45. NC + PREP + NC + ADJ + ADJ
46. NC + PREP + NC + ADJ + PREP + NC
47. NC + PREP + NC + ADJ + PREP + NC + ADJ
48. NC + PREP + NC + ADJ + PREP + NC + PREP + NC
49. NC + PREP + NC + ADJ + PREP + NC + PREP + NC + ADJ
50. NC + PREP + NC + ADV + ADJ
51. NC + PREP + NC + PREP + ADJ
52. NC + PREP + NC + PREP + ADJ + NC
53. NC + PREP + NC + PREP + ADJ + NC + PREP + NC
54. NC + PREP + NC + PREP + NC
55. NC + PREP + NC + PREP + NC + ADJ
56. NC + PREP + NC + PREP + NC + ADV + ADJ
57. NC + PREP + NC + PREP + NC + PREP + NC
58. NC + PREP + NC + PREP + NC + PREP + NC + ADJ
59. NC + PREP + NC + PREP + NC + PREP + NC + PREP + NC
60. NC + PREP + NC + PREP + NC + VLadj
61. NC + PREP + NC + VLadj
62. NC + VLadj
63. NC + VLadj + ADJ
64. NC + VLadj + ADJ + PREP + NC
65. NC + VLadj + PREP + NC
66. NC + VLadj + PREP + NC + ADJ
67. NC + VLadj + PREP + NC + ADJ + ADJ
68. NC + VLadj + PREP + NC + ADJ + PREP + NC
69. NC + VLadj + PREP + NC + PREP + NC
70. NC + VLadj + PREP + NC + PREP + NC + ADJ

Annex 2. Paper prototype

Task 1:

Please, read carefully. There is no hurry!;)

Task 1: Use this service to find an English translation for the Spanish term “*regulador de voltaje*”.

Figure 55. Task 1. Instruction page.

Choose your language pair:

English → Spanish

Spanish → English

Term to translate:

Figure 56. Task 1. Screen 1.

No exact match found for “*regulador de voltaje*”. Please select the related/relevant terms below:
*Required

regulador luz 12v

regulador caja de muerto

regulador de voltaje de alternador

regulador de generador

Are these images of your object? (Check all that apply)
* Optional







Figure 57. Task 1. Screen 2.

Are these images of your object? (Check all that apply)

* Optional







Check the terms you consider candidate translations (Check all that apply):
*Required

fuel pressure regulator
 adjustable fuel regulator
 distributor vacuum regulator

electric fuel pump regulator
 bonnet rear rubber seal
 alternator voltage regulator

remote fuel pressure regulator
 air filter element
 injection pressure regulator

Submit

Figure 58. Task 1. Screen 3.

Your Terms:

<p>Subject field: automotive engineering</p> <p>eng: alternator voltage regulator term source: http://www.globalsources.com/manufacturers/Alternator-Voltage-Regulator.html context: IN2205E Agriculture Equipment AVR 12V Automatic Alternator Voltage Regulator Automobile Accessories</p> <p>spa: regulador de voltaje para alternador term source: http://www.wiju.es/motor/alternador-renault-9-y-su-regulador.html context: REGULADOR VOLTAJE DE ALTERNADOR PARA RENAULT 14, NUEVO Y CON SU CAJA FEMSA</p> <p>spa: regulador de voltaje term source: user1 context:</p>	<p>Image:</p> 
--	---

New search

Figure 59. Task 1. Screen 4.

Task 2:

Please, read carefully. There is no hurry!;)

Task 2: Use this service to find an English translation for the Spanish term “*estabilizador de corriente*”.

Figure 60. Task 2. Screen 1.

Choose your language pair:

English → Spanish
 Spanish → English

Term to translate:

Submit

Figure 61. Task 2. Screen 2.

No term match for “*estabilizador de corriente*”. Try uploading your own image of the object (white background works better):

Browse...

Figure 62. Task 2. Screen 3.

No term match for “*estabilizador de corriente*”. Try uploading your own image of the object (white background works better):



Figure 63. Task 2. Screen 4.



Figure 64. Task 2. Screen 5.

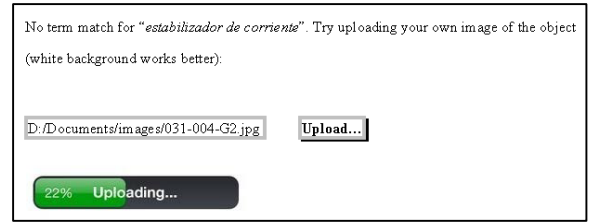


Figure 65. Task 2. Screen 6.

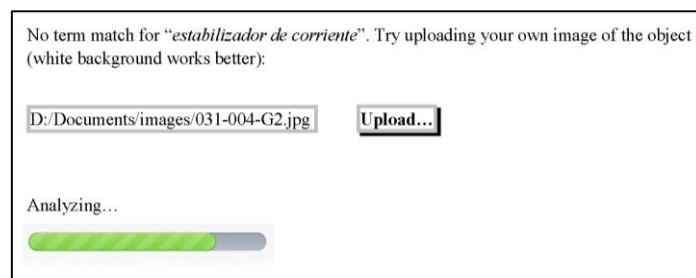





Figure 66. Task 2. Screen 7.

<p>Are these images similar to yours? (Check all that apply):</p> <p>* Optional</p> <p><input type="checkbox"/> </p> <p><input type="checkbox"/> </p> <p><input type="checkbox"/> </p>	<p>Check the terms you consider candidate translations (Check all that apply):</p> <p>*Required</p> <p><input type="checkbox"/> fuel pressure regulator</p> <p><input type="checkbox"/> adjustable fuel regulator</p> <p><input type="checkbox"/> distributor vaccum regulator</p> <p><input type="checkbox"/> electric fuel pump regulator</p> <p><input type="checkbox"/> alternator voltage regulator</p> <p><input type="checkbox"/> bonnet rear rubber seal</p> <p><input type="checkbox"/> remote fuel pressure regulator</p> <p><input type="checkbox"/> air filter element</p> <p><input type="checkbox"/> injection pressure regulator</p>
---	---

Submit

Figure 67. Task 2. Screen 8.

Your Terms:

<p>Subject field: automotive engineering</p> <p>eng: alternator voltage regulator</p> <p>term source: http://www.globalsources.com/manufacturers/Alternator-Voltage-Regulator.html</p> <p>context: IN220SE Agriculture Equipment AVR 12V Automatic Alternator Voltage Regulator Automobile Accessories</p> <p>spa: regulador de voltaje para alternador</p> <p>term source: http://www.wiiu.es/motor/alternador-renault-9-y-su-regulador.html</p> <p>context: REGULADOR VOLTAJE DE ALTERNADOR PARA RENAULT 14, NUEVO Y CON SU CAJA FEMSA</p> <p>spa: regulador de voltaje</p> <p>term source: user1</p> <p>context:</p>	<p>Image:</p> 
--	---

Figure 68. Task 2. Screen 9.