

Semi-automatic Generation of Subcategorization Frames for Spanish Verbs Using Ontologies and Verbs Functional Class

Rodolfo A. Pazos R.

Instituto Tecnológico de Cd. Madero/División de Estudios de Posgrado en Investigación, Cd. Madero, Mexico
 Email: r_pazos_r@yahoo.com.mx

José A. Martínez F., Javier González B., María L. Morales-Rodríguez, and Alberto Castro H.
 Instituto Tecnológico de Cd. Madero/División de estudios de posgrado en Investigación, Cd. Madero, Mexico
 Email: {jose.mtz, lmoralesrdz}@gmail.com, {jjgonzalezbarbosa, a_castro_h}@hotmail.com

Abstract— This work deals with the semi-automatic generation of subcategorization frames (SCFs) of Spanish verbs; specifically, given a set of verbs in Spanish and their respective sense, their SCFs are obtained. The acquisition of SCFs in Spanish has been approached in different works: in some the frames are generated manually, while in others they are obtained semi-automatically from a tagged corpus; unfortunately in this case, the results depend on the characteristics of the texts used. The method proposed in this document combines an ontology-based approach (through lexical relations of verbs) and linguistic knowledge (functional class of verbs). The relations among base verbs and other verbs were obtained from the Spanish WordNet ontology, which contains lexical relations among words. Also, the existing relation between the SCF and the functional class of verbs was used to generate the SCFs. In order to evaluate the method, the SCFs for 44 base verbs were generated manually, from which 239 SCFs were automatically generated and validated, yielding an accuracy of 89.38%.

Index Terms— Subcategorization Frames, Spanish verbs, Meaning-Text Theory, lexical relations.

I. INTRODUCTION

Most Natural Language processing applications use syntactic parsing. There exist several theories for parsing. The most used ones are constituent parsing and dependency parsing. Some of the most important dependency parsing techniques are based on the Meaning-Text Theory (MTT). The use of MTT allows to describe Spanish characteristics, as the establishment of relations among syntactic and semantic valencies.

Verb subcategorization frames (SCF) constitute one of the most important elements of lexical/grammatical knowledge for efficient and reliable parsing [1]. The implementation of an MTT parser requires creating a base of linguistic knowledge that contains the SCFs of verbs that allow to know how words are related in a sentence.

At the Centro Nacional de Investigación y Desarrollo Tecnológico (Mexico) a syntactic parser for a natural

language interface was developed, which uses an MTT dependency parser. This parser uses a base of SCFs for verbs, which does not contain enough linguistic information for parsing.

The parser uses a base of SCFs as linguistic knowledge [2] for filtering syntactic structures and discarding invalid ones. The database of the linguistic knowledge for the SCFs is implemented in PostgreSQL 7.2.1 and consists of five tables: Verbs, Classes, Sentences, SCFs and Combinations. Figure 1 shows the tables of the database of SCFs and their relations

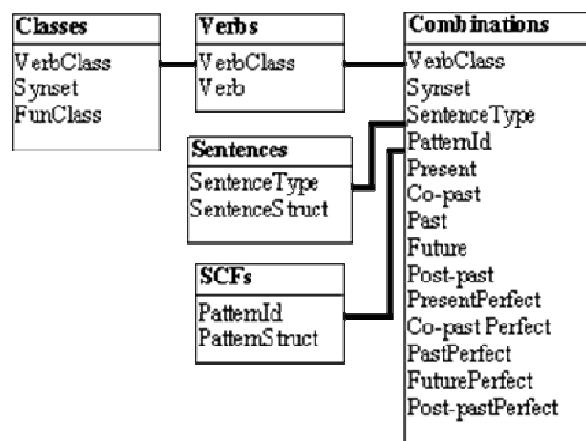


Figure 1. Schema of the database of subcategorization frames.

II. PREVIOUS WORK

In [1] a statistical-based mechanism for automatically acquiring SCFs and their frequencies based on tagged corpora is described. The authors showed that by combining syntactic and statistical analysis, SCF frequencies can be estimated with large accuracy. Unfortunately, the frequencies of just six SCFs were measured, but they claim that their method is general and can be extended to much more SCFs.

A cooperative machine learning system (ASIUM) is presented in [3], which is able to acquire SCFs with restrictions of selection and ontologies for specific domains from syntactically parsed technical texts in natural language. Experiments on corpora of cooking recipes in French, and patents in English, have shown the applicability of the method to texts in restricted and technical domains and the usefulness of the cooperative approach for such knowledge acquisition. Unfortunately, no quantitative results on the effectiveness of ASIUM are provided because "evaluation of the unsupervised learned knowledge quality is a very difficult problem for which we have currently no solution" [3].

In [4] a linguistic-based work is presented that proposes a method for adding semantic annotation to verb predicate arguments for a lexicon called Penn English TreeBank. The work relies considerably on recent works on linguistics about word classification that has more semantic orientation, such as the Levin classes and WordNet. The document mentions that 1000 SCFs have been manually generated, and hints at the possibility of automatic frame generation; for example, it mentions that a large number of SCFs for repetition or negated actions (verbs) can be generated (e.g., state/restate, load/unload).

In [5] a linguistic-based work is presented, which proposes an extension of 57 new classes to Levin's verb taxonomy. An experiment was carried out in the context of automatic SCF corpus-based acquisition for determining the usefulness of the new classes proposed, which yielded an accuracy of 71.0% (the percentage of SCF types proposed by the system which are correct). Though several classifications are currently available for English verbs, they are restricted to certain class types and many of them have few instances in each class. The document mentions that some experiments recently reported indicate that it should be possible in the future to automatically supplement existing classifications with original/new verb classes and member verbs from corpus data.

In [6] a probabilistic classifier (based on C5.0) is described, which can automatically classify a set of verbs into argument-structure classes with a reasonable error rate (33.4%). The method exploited the distributions of selected features from the local context of the verb, which were extracted from a 23-million word corpus from the Wall Street Journal.

In [7] a method is explained, which uses unsupervised learning to learn equivalence classes for verbs in Spanish. Specifically, the approach consists of using the Expectation Maximization algorithm for clustering verbs according to their contexts of occurrence, which would permit extrapolating the behavior of known verbs to unknown ones. The authors claim having found a good clustering solution that distinguishes verbs with clearly different SCFs. The effectiveness of this approach ranges from 66 to 83%.

As can be seen from this survey on SCFs, some works do not deal with automatic generation of SCFs or do not provide quantitative results on the effectiveness of the proposed approaches. Finally, as will be seen in Sections

"VI. Experimentation" and "VII. Conclusions", the method proposed in this paper compares favorably with respect to other approaches.

III. SUBCATEGORIZATION FRAMES

Subcategorization identifies the arguments that are grouped by one sentence element, which can be a verb, a noun or an adjective [8].

The linguistic knowledge about some part of a sentence, is known in dependency parsing techniques as *valency*.

Valency is known in chemistry as the capacity of a chemical element that makes possible the combination of a fixed number of atoms with some other element. For example, oxygen is divalent because it has the capacity to be linked to two atoms.

Likewise, some types of words have the capacity of demanding arguments. If their demand is satisfied, completely intelligible sentences and phrases are derived [9]. Therefore, valency is the number of arguments that a word requires.

For example, in (1) the verb *ate* is the main part of the sentence and governs two phrases: *your brother's girl friend* and *salad*.

Our brother's girl friend ate salad. (1)

Sentence (1) can be extended as in (2); however, the valency of verb *eat* requires only two arguments (called actants in MTT), which consist of someone (*your brother's girl friend*) and something (*salad*). With only these actants the sentence is completely intelligible.

Your brother's girl friend ate salad yesterday in my home. (2)

The additional complements in the previous extended sentence that specify time (*yesterday*) and the place of action (*in my home*) are not governed by the valency. Both complements are optional and are called adjunct. The denominations of adjuncts (i.e., temporal, causal, modal and conditional adjuncts) follow the terminology of the traditional grammar. Generally, all the relative and adverbial clauses belong to the class of adjuncts [9].

Additionally, the syntactic valency is also considered, which is defined as the syntactic materialization of the semantic valency. In the previous example, the materialization of someone who eats is a noun syntagm, and the materialization of what is eaten is, also, a noun syntagm.

It is worth mentioning that, although the semantic actants are required by the verb, their syntactic materialization might not be compulsive. For example, in (3) for the verb *buy*, syntactically and semantically there exist four actants: who buys (*John*), what is bought (*a car*), whom it is bought from (*Louis*) and how much is paid (*9,000 dollars*).

John bought a car from Louis for 9,000 dollars. (3)

However, in (4) the same verb *buy* is used, but only two actants are syntactically materialized: who buys (*Annie*) and what is bought (*an ice cream*). Though,

semantically, it is implied that the ice cream was bought from someone and it had a purchase price.

Annie bought an ice cream. (4)

Finally, it is not necessary that all the semantic valencies have a syntactic materialization (syntactic valency) for a sentence to be intelligible.

Subcategorization frames specify the category of the main anchor, the number of arguments, each argument's category and position with respect to the anchor, and other information [10]. Figure 2 shows an example of the structure of the SCF for verb *inhalar* (inhale).

```
00004020
inhalar1 (inhale)
(Medicine) X draw in air
X=1
1. SN
```

C₁₁: SN indicates a nominal syntagm with animate noun.

Combinations	
C..	{Un buen hombre} inhala [. exhala y acomoda sus ideas] {A good man} inhales [. exhales and accommodates his ideas]

Figure 2. Subcategorization frame of the verb *inhalar* (inhale).

III. GENERAL DESCRIPTION OF THE METHOD

The method proposed for generating SCFs consists of two phases:

- Generate manually the SCFs of a group of verbs/sense denominated *base verbs/sense*. The steps carried out for generating manually the SCFs of the verbs/sense were the following:
 - Select a verb/sense from the base verbs/sense set.
 - Generate manually the SCF of the selected verb/sense according to its meaning.
 - Create the syntactic structure of a sentence with respect to the verb, according to the examples presented in the gloss (if it exists).
 - Check the SCF of the verb/sense against sentences in the corpus published by the Real Academia Española [11].
 - Verify the SCFs by an expert in linguistics.
- Using the SCFs of the base verbs/sense, automatically obtain the SCFs of other verbs/sense with related meaning. There exist two alternatives for this step:
 - Assignment. When the functional class of the verb/sense being processed has the same functional class as some of its hypernym verb/senses, the verb/sense in question inherits the SCF of the matching hypernym.
 - Elaboration. When the functional class of the verb/sense being processed does not match the functional class of any of its hypernym verb/senses, its SCF is obtained from the SCF of one of its hypernyms and modifying this SCF (by appending

or eliminating elements) so as to make it compatible with the functional class of the verb/sense being processed.

For developing a database filling program based on the preceding method, two programs were designed: the Automatic Overall Filling Sub-module, which uses the base verbs/sense to fill the base of SCFs automatically for all the verbs; and the Automatic Levelwise Filling Sub-module, which uses a file that contains a verb/sense that considers as base, and obtains the SCFs of the hyponyms of the verb/sense. Both algorithms use the Verb Verification Submodule, which verifies that the verb/sense exists in the Spanish WordNet ontology. If it exists, it is processed by the Querying, Assignment and Frame Elaboration Sub-module, for generating its SCF. (More details can be found at [12].)

IV. LEXICAL RELATIONS

In order to relate the base verbs/sense with other verbs, it was necessary to obtain the lexical relations that exist among them. To this end the hyponymy relation (has_hyponym) contained in the Spanish WordNet ontology 1.0 was used. This relation allows to connect synsets of the ontology in a hierarchy, which is important to obtain the SCFs.

For example, Figure 3 depicts the semantic relation of verbal hyponymy of the verb/sense *respirar*₁ (breathe). We hypothesize that if one knows the SCF of the verb/sense *respirar*₁ (breathe), one can generate the SCF of the verb/sense *espirar*₁, *exhalar*₁ (breathe out, exhale), because this verb/sense is hyponym of *respirar*₁. Also, if one knows the SCF of this verb/sense, one can generate the SCF of the verb/sense *soplar*₁ (blow) for the same reason. And this one, in turn, will serve to generate the SCF of the verb/sense *resoplar*₂ (puff).

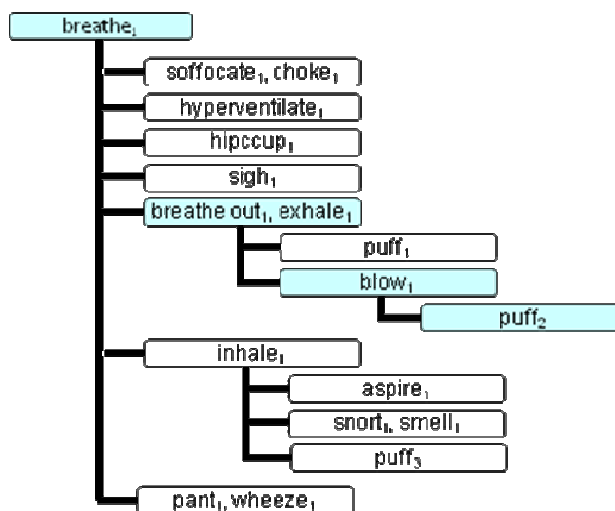


Figure 3. Example of lexical relation of verbal hyponymy in Spanish WordNet 1.0.

V. SUBCATEGORIZATION FRAMES OF BASE VERBS/SENSE

In Spanish WordNet there exist 366 synsets in the highest level of the ontology. It is important to mention that each synset contains one or more verbs/sense. These verbs/sense were denominated base verbs/sense or level-1 verbs/sense. Since the manual generation of SCFs for the 366 level-1 verbs/sense is an arduous task that requires a lot of time, we worked with a sample of 44 level-1 verbs/sense, which were generated manually.

It is important to mention that in this work, we consider that the inclusion of the Spanish preposition *a* (to) in a direct complement, which defines the animacy of the complement, depends on the semantics of the noun that it precedes. Therefore, the verb does not determine if its direct complement includes preposition *a* (to) [13].

VI. EXPERIMENTATION

A. Test Cases

The sample of 44 base verbs/sense that was used for the tests, belong to level 1 of the Spanish WordNet ontology. From these, the first test cases were obtained, which correspond to the verbs/sense of level-2 synsets (hyponyms of level-1 synsets), for which their SCFs were generated (derived or assigned). These SCFs served as a basis to obtain the SCFs of the verbs/sense of level-3 synsets (hyponyms of level-2 synsets). This process can be continued up to the last level.

It is necessary to mention that, each time that the generation of the SCFs of the verb/sense of some level begins, it has to be verified that the SCFs of their hypernym synsets are correct.

As can be seen in Figure 4, the test cases correspond to the shaded area of Spanish WordNet.

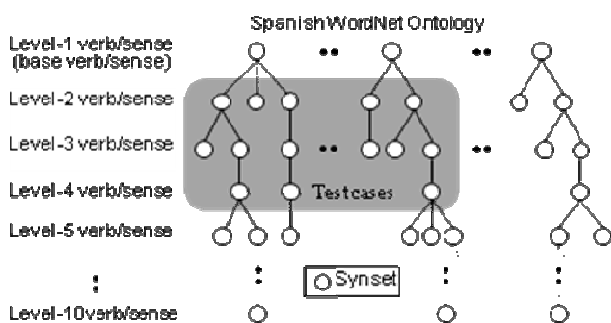


Figure 4. Test cases

B. Experimentation

The experimentation consisted of three tests; and in order to verify that the 195 SCFs were correctly obtained by the SCFs generation method, these were previously generated by hand. The experiments conducted were the following:

1. The first experiment consisted of elaborating or assigning the SCFs for 100 level-2 verbs/sense that are hyponyms of level-1 verbs/sense.
2. The second experiment consisted of elaborating or assigning the SCFs for 60 level-3 verbs/sense that

are hyponyms of level-2 verbs/sense. The SCFs for level-2 verbs/sense were previously checked, and corrected for those verbs/sense whose SCFs were incorrectly generated in the first experiment.

3. The third experiment was similar to the second, except that in this case SCFs were generated for 35 level-4 verbs/sense. Likewise, this process continues up to the last level.

C. Analysis of Results

In order to assess the effectiveness of the generation method of SCFs, the percentage of correct results was calculated for the SCFs generated.

The percentage of success (PS) of a verb/sense was obtained comparing each one of the possible combinations of the correct SCF with the possible combinations of the corresponding obtained SCF, and using the following formula:

$$PS = \frac{\text{No. of correct combinations generated}}{\text{No. of correct combinations possible}} \times 100$$

In the preceding formula the number of correct combinations generated (NCCG) is divided by the number of correct combinations possible (NCCP). Additionally, the number of incorrect combinations generated (NICG) and the number of over-generated combinations (NOC) were calculated.

Table 1 shows the obtained results, which includes the average of PA and the percentage of NICG for each experiment, and the overall averages for all the experiments.

Table 1. Success percentage from the experiments

Experiment	Average of PS	Percentage of NOC
1	81.00%	23.53%
2	90.00%	0.00%
3	97.14%	0.00%
Overall average	89.38%	7.84%

The experiments yielded some errors in the derivation/assignment of the SCFs. The errors found are explained hereupon:

- Errors generated in the elaboration mode:
 - Verb/sense without necessary complements. This error occurred when generating a SCF that does not accept Prepositional Complements (PC) or Attributes (Atr) that are necessary for some verbs. For example, the SCF generated for the verb *convertirse*₁ (become) allows the incorrect sentence (5).

[Subj] *La plática* (The chat) [v] *se convirtió* (became). (5)

Whereas the correct SCF must have a prepositional complement as shown in sentence (6).

[Subj] *La plática* (The chat) [v] *se convirtió* (6)

(became)] [_{PC} *en un desastre* (a disaster)].

The SCF of the verb/sense *convertirse*₁ (become) was obtained from the SCF [_{Subj(NS)}] V for the verb/sense *cambiar*₂ (change). This SCF does not include a PC (*en NS*); and therefore, the SCF elaborated for the verb/sense *convertirse*₁ (become) is incorrect.

- Transitive verb/sense with incorrect complement. This error occurred when generating the following SCF for transitive verbs: [_{Subj(S)}] V DC(*a N|Pro*). Some complement(s) is(are) added to this SCF as Circumstantial Complement (CC), which is obtained from its hypernym. Such kind of complement is incorrect for some transitive verbs. For example, the SCF generated for the verb/sense *hacer*₁ allows the incorrect sentence (7).

[_{Subj} *Samuel*] [_V *hizo* (made)] [_{DC} *a su novia* (his girlfriend)] [_{CC} *bien* (good)]. (7)

Whereas the correct SCF accepts an attribute as in the following sentence.

[_{Subj} *Samuel*] [_V *hizo* (made)] [_{DC} *a su novia* (his girlfriend)] [_{Attr} *feliz* (happy)]. (8)

The SCF of the verb/sense *hacer*₁ (make) was obtained from the SCF [_{Subj (NS)}] Pro-V CC(Adv) for the verb/sense *cambiar*₂ (change). This SCF includes the DC(Adv) and it doesn't include Atr(Adj); and therefore, the SCF that was elaborated for the verb/sense *hacer*₁ (make) is incorrect.

- Errors generated in the assignment mode:

- Transitive verb/sense with optional Direct Complement. This error occurred when copying the SCF to a transitive verb, whose DC is optional. For example, the SCF generated for the verb *ensanchar*₄ (widen) is incorrect because it allows the absence of a DC as in the following sentence (9):

[_{Subj} *Los trabajadores* (The workers)] [_V *ensancharon* (widened)] (9)

when the Direct Complement must be mandatory; for example, sentence (10).

[_{Subj} *Los trabajadores* (The workers)] [_V *ensancharon* (widened)] [_{CD} *la carretera* (the highway)]. (10)

The SCF of the verb/sense *ensanchar*₄ was obtained from the SCF [_{Subj(NS)}] V DC(NS) for the verb/sense *cambiar*₃ (change). This SCF includes an optional DC; therefore, the SCF assigned to the verb/sense *ensanchar*₄ (widen) is incorrect.

- Verb/sense with incorrect complements. This error occurred when copying the SCF from a hypernym verb to a hyponym verb, whose complement (indirect, prepositional or attribute) is incorrect for

the last verb. An example of this error occurs with the verb/sense *fanfarronear*₁ (brag) in sentence (11).

[_{Subj} *Juan* (John)] [_V *fanfarronea* (brags)] [_{CC} *mal* (wrong)]. (11)

Whereas the correct form of the previous sentence is (12).

[_{Subj} *Juan* (John)] [_V *fanfarronea* (brags)]. (12)

The SCF of the verb/sense *fanfarronear*₁ (brag) was obtained from the SCF [_{Subj(NS)}] V DC(Adv) for the verb/sense *actuar*₁ (act). This SCF includes the DC(Adv); and therefore, the SCF that is assigned to the verb/sense *fanfarronear*₁ is incorrect.

- Verb/sense without necessary complements. This error occurred when copying a SCF to a verb, which does not include Prepositional Complements (PC) or Circumstantial Complements (CC), which are necessary for some verbs. For example, the SCF generated for the verb/sense *comportarse*₂ (behave) does not allow the correct sentence (13).

[_{Suj} *Luis*] [_V *se comportó* (behaved)] [_{CC} *como loco* (as a lunatic)]. (13)

The SCF of the verb/sense *comportarse*₂ (behave) was obtained from the SCF [_{Subj(NS)}] V CC(Adv) for the verb/sense *actuar*₁ (act). This SCF does not include the CC(*como NS|Adv*); and therefore, the SCF that is assigned to the verb/sense *comportarse*₂ (behave) is incorrect.

VII. CONCLUSIONS

A total of 239 SCFs of Spanish verbs were generated manually and validated. A program was developed based on the proposed method for SCF generation. The results obtained on the generation of the SCFs of verbs/sense in three levels (level 2, level 3 and level 4) of the Spanish WordNet ontology yielded an average success of 89.38% and an average percentage of 7.84% for over-generated combinations, which compares favorably with the 71.0% reported in [5] and the 83% reported in [7].

The experiments revealed that the percentage of success of the SCFs generated improves as the level of depth in the ontology increases. A possible explanation for this behavior is that the meaning of these verbs makes it possible that less verb valencies are required.

REFERENCES

- [1] Ushioda, A., Evans, D.A., Gibson, T., Waibel, A.: The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora, SIGLEX ACL Workshop on *The Acquisition of Lexical Knowledge from Text*, Columbus, Ohio, pp. 95-106 (1993).
- [2] Cervantes A.: Diseño e Implementación de un Analizador Sintáctico para las Oraciones en Español Usando el Método de Dependencias. MS thesis. Centro Nacional de Investigación y Desarrollo Tecnológico (2005).

- [3] Faure, D., Nedellec, C.: Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM. In Proc. of the 11th European Workshop on Knowledge Acquisition, Modeling and Management, pp. 329-334 (1999).
- [4] Kingsbury, P., Marcus, M., Palmer, M.: Adding Semantic Annotation to the Penn Tree-Bank. In Proc. of the Human Language Technology Conference (HLT), San Diego, CA (2002).
- [5] Korhonen, A.: Assigning Verbs to Semantic Classes via WordNet. In Proc. of the SemaNet'02: Building and Using Semantic Networks. Taipei, Taiwan pp. 1-7 (2002).
- [6] Sarkar, A., Tripasai, W.: Learning Verb Argument Structure from Minimally Annotated Corpora. In Proc. of the Int. Conf. on Computational Linguistics. Taipei, Taiwan pp. 1-8 (2002).
- [7] Castellón, I., Alemany, L.A., Tincheva, N.T.: A Procedure to Automatically Enrich Verbal Lexica with Subcategorization Frames. *Revista Iberoamericana de Inteligencia Artificial*, Vol. 12, N°. 37. pp. 45-53 (2008).
- [8] Flores D., "Cómo Crear un Lenguaje", http://www.pueblacity.com/ego-pdf/sp/lng/como/como_verbos.html.
- [9] Dobrenov-Major M., "Reading Comprehension Enhancement in Foreign Language Learners with a Verb Valency Based Reading-Strategy", Applied Linguistics Association of Australia (1998).
- [10] *XTAG Project*, "Subcategorization Frames", <http://www.cis.upenn.edu/~xtag/tech-report/node248.html> (2008).
- [11] REAL ACADEMIA ESPAÑOLA, Banco de datos (CREA), Corpus de referencia del español actual, <http://www.rae.es> (2006).
- [12] Castro A.: Llenado de una Base de Patrones para la Valencia de los Verbos del Idioma Español". M.S. thesis, Instituto Tecnológico de Cd. Madero (2006).
- [13] Galicia S.: Análisis Sintáctico Conducido por un Diccionario de Patrones de Manejo Sintáctico para Lenguaje Español. PhD dissertation. Centro de Investigación en Computación, Instituto Politécnico Nacional (2000).

Rodolfo A. Pazos R. was born in Tampico, Mexico in 1951. He received a BS in Electronics Engineering from the Instituto Politécnico Nacional (Mexico) in 1973, a MS degree in Electrical Engineering from CINVESTAV (Mexico) in 1976 and a PhD in Computer Science from UCLA (USA) in 1983.

He is currently working as full professor at the Instituto Tecnológico de Cd. Madero, Mexico. He has also worked as professor at CENIDET (México), IIE (Mexico) and CIDET (Mexico).

Dr. Pazos has been distinguished as a Level II member of the Sistema Nacional de Investigadores (CONACYT, Mexico).

José A. Martínez F. is full professor at the Instituto Tecnológico de Cd. Madero, Mexico. He received a Ph.D. degree in Computer Science from CENIDET, Mexico in 2006. He has conducted research in database systems and natural language interfaces. Dr. Martínez has been distinguished as member of the Sistema Nacional de Investigadores (CONACYT, Mexico).

Javier González B. is full professor at the Instituto Tecnológico de Cd. Madero, Mexico. He received a Ph.D. degree in Computer Science from CENIDET, Mexico in 2005. He has conducted research in natural language interfaces. Dr. González has been distinguished as a Level I member of the Sistema Nacional de Investigadores (CONACYT, Mexico).

María L. Morales-Rodríguez. is professor at the Instituto Tecnológico de Cd. Madero, Mexico. She received a Ph.D. degree in Informatics from Université Paul Sabatier, France in 2007. She has conducted research in virtual characters and emotional interfaces.

Alberto Castro H. is professor at the Universidad Tecnológica de Altamira, Mexico. He received a MS degree in Computer Science from the Instituto Tecnológico de Cd. Madero, Mexico in 2006. He has conducted research in natural language interfaces.