# Extracting Human Spanish Nouns*

Sofia N. Galicia-Haro[1] and Alexander F. Gelbukh[2]

[1] Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico, D. F.
sngh@fciencias.unam.mx
[2] Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico, D. F.
gelbukh@gelbukh.com

**Abstract.** In this article we present a simple method to extract Spanish nouns with the linguistic property of "human" animacy. We describe a non-supervised method based on lexical patterns and on a person name list enlarged from a collection of newspaper texts. Results were obtained from the Web filters and estimation methods are proposed to validate them.

**Keywords:** Animacy, human mark, Spanish nouns, non supervised learning.

## 1 Introduction

In general, the animacy mark distinguishes living entities from non-living ones. But animacy might be considered as a range that goes from "human" consideration to "inanimate objects". For example, [17] analyze the codification of animacy in English. They distinct three categories: human beings, other animates and inanimates.

Animacy is an important category in linguistic analysis. Animacy has effects in grammar, in word order, in sentence production, etc. For example, in Spanish, reference to a direct object that is a human being makes the inclusion of the "a" preposition obligatory [1]; John Myhill discusses how in Chorti, a Mayan language that exhibits a strong tendency to VO order, animate subjects appear more in a preverbal position than inanimate subjects [16]; in English, control verbs in irregular past participles (e.g.: eaten, shaken) prefer animate subjects in active sentences [2].

For these reasons, in natural language processing, automatic animacy identification is important. Researchers have analyzed its importance in generation and translation [17], in parsing [12], in anaphora resolution [11], etc. Nevertheless, the animacy mark is not found systematically in common resources. Seeking in a Spanish dictionary such as DRAE[1] or MOLINER[2] the *coronel* 'colonel' noun, we find the same main description: *Jefe militar que manda un regimiento* 'Military chief that commands a regiment', but they do not mark explicitly its human condition, because it is clear for a human reader. Other nouns such as *basquetbolista* 'basketball player' (absent in DRAE) or *narcotraficante* 'drug dealer' (appearing as an adjective in DRAE), are nouns [+human]. So, it is not possible to extract this mark from such resources.

* Work done with partial support of Mexican Government (CONACyT, SNI, CGEPI-IPN).

[1] http://buscon.rae.es/draeI/

[2] http://www.diclib.com/cgi-bin/d1.cgi?l=es&base=moliner&page=showindex

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 75–83, 2010.

Researchers have conducted work to automatically identify animacy. For example, in [10], the authors present a method for English nouns using WordNet and machine learning techniques, and their evaluation results show that animate entities are more difficult to identify than inanimate. In [8], the authors develop a simple approach to discover gender and animacy knowledge. They automatically discover a large knowledge base of gender and animacy properties for noun phrases, with animacy based on a pronoun pattern for *who*.

The aim of our work is to annotate "human" animacy in Spanish texts. In this work, we obtain a list of "human" animacy nouns automatically. This paper is structured as follows. First we describe some characteristics of animacy in Spanish constructions. In Section 3 we present the proposed technique to obtain human animate nouns using instances like proper names. In Section 4 we present the results when applying this technique in Web-scale n-grams. Finally, in Section 5 we present the conclusions.

## 2 Animate Nouns in Spanish

Spanish, like other Indo-European languages, has grammatical gender: nouns are partitioned into sets which, in general, contrast distinctions of sex or animateness. Spanish has two genders (feminine, masculine), German has three genders (neuter, in addition). Identification of human animated nouns would be much easier if language could assign a precise gender for people. But languages like Spanish assign genders such as masculine and feminine to inanimate. Nouns denoting people, assigned to masculine or feminine gender according to sex[3], are a minority [6]. The "exceptions" (non-sexed objects assigned to either of those two genders) are the majority, thus making the semantic association a rather useless predictor for the gender of a noun.

Animate nouns in Spanish can be identified in different contexts which can be divided into several types:

– As a direct object

The animate direct object requires a preposition to be linked to the verb, becoming a prepositional complement. The so-called Prepositional Direct Complement is a topic much discussed in Spanish grammar. It is a linguistic phenomenon present in many languages [14], known as Differential Object Marking (DOM). In such languages, direct objects can be divided into two different classes and only one class receives a mark. In Spanish the mark corresponds to the preposition "a" when the noun is animate. For example:

*Veo esa casa*. 'I can see that house' (inanimate direct object)
*Veo a esa niña*. 'I can see that girl' (animate direct object)

This form is usual in standard Spanish. Nevertheless, in several dialects of Spanish, especially in Latin America, the preposition "a" precedes direct objects which are not animate if they are definitive and specific [15]. For example:

*Vio a las sierras*. 'He saw the saws' (Puerto Rican Spanish)
*Cosecharon al maíz*. 'They harvested the corn' (Argentinian Spanish)

[3] Author use the word "sex" to refer to biological gender, reserving "gender" for the grammatical category.

– According to the verb

Verbs select subjects according to their animate or inanimate condition. For example: *la madera cruje* 'the wood creaks', **Juan cruje* 'John creaks', although in some cases, the metaphorical sense could change a non-grammatical condition to a correct object selection. For example: ... *y crujo como sal que se derrite* '... and I rattle as salt that melts'.

In [3] the authors analyze the so-called verbs of temporal expression: *durar* 'to last' and *tardar* 'to take time'. Moliner's definition of the verb "to last" in its first meaning is "To be a thing happening, existing, working, etc., the time that expresses itself". Both are intransitive verbs and they may have a subject [+animate]. Thus, with respect to the verb *tardar*, in a phrase such as a *Julia tarda media hora en hacer un ejercicio* 'Julia takes half an hour doing an exercise', the verb makes reference to the time a subject [+ animate] employs in carrying out certain activity.

Nevertheless, such a syntactic structure (subject + verb + temporary complement + supplement) is impossible to realize with the verb "to last", which, in these cases, does not admit the feature [+animate]: the phrase is ungrammatical: **Julia dura treinta minutos en hacer un ejercicio* 'Julia lasts thirty minutes doing an exercise'. The authors indicate that, to be able to use the verb *durar* with a subject [+animate], it must be used in cases in which its meaning acquires other submeanings, with certain nuances that distinguish them from the first meaning: in these cases the verb *durar* means: to endure, to continue, to keep up.

In [14], the authors analyze different classes of verbs. They consider as their main sources of analysis the Bible and the Corpus del Español (s. XII to XIX). Their analysis confirms the hypothesis that the class of verb is the principal parameter for DOM in Spanish. Another important conclusion is that the direct object mark is determined by parameters in a multi-dimensional space.

– Pronouns

With pronouns, there is a tendency to use *le* like the pronoun of a direct object, although it is usually an indirect object pronoun or accusative pronoun, meaning 'to him/her', at the expense of the pronouns that are the accusative pronouns, when the referent is animate. DRAE[4] exposes that among other classes the so-called verbs of *psychic condition,* those that designate processes that affect encouragement or produce actions or emotive reactions, like affecting, scaring, amazing, convincing, etc., depending on different factors, admit the use of the accusative pronouns: *lo*(s), *la*(s), and the dative pronouns: *le*(s). The selection of one or other of these depends basically on whether the subject is or is not an active agent of the action and on the grade of volition that he has or assumes with regard to the action designated by the verb. With animate subjects this alternation can happen also if the action denoted by the verb is realized voluntarily or not for the subject. For example: *Su padre, que se había disfrazado, lo asustó* 'His father, who had disguised himself, scared him' (he gave him a fright intentionally), *Su padre, que se había disfrazado, le* asustó 'His father, who had disguised himself, scared him' (the fright is involuntary, the cause is the fact of going in disguise).

[4] `http://buscon.rae.es/dpdI/SrvltGUIBusDPD?lema=le%EDsmo`

– In noun apposition

An apposition is a construction of two close grammatical elements, the second of which specifies the first. In the juxtaposition of one noun to the other by means of the apposition, compounds are formed by two nouns which are written together (*compraventa* 'buying and selling') or separately (*compra venta*). In the case of animate nouns, a common apposition is that of a proper noun to another generic. This apposition specifies a personal characteristic (*lawyer Juan Torres, your brother Juan*). To create an information repository that helps to answer a question, [5] consider patterns to extract highly precise relationship information. The most productive patterns considered are two syntactic constructions that often indicate the relationship concept–instance, common noun-proper name and appositions, such as *President George Bush*.

## 3 Instances of Human Animate Nouns

The annotation of animacy is not standard in corpora or treebanks. Studies in the corpus on animacy, for example [10], have used data with manual annotations. Those annotations differ according to the considered scheme, with diverse granularity of categories. We consider in this work only one distinction between Human and Non Human, abbreviated as [+H] and [-H].

As [13] pointed out, the attributes of a given class can be derived by extracting and inspecting the attributes of individual instances from that class. For example, the attributes of the class Car are extracted by inspecting attributes extracted for Chevrolet Corvette, Toyota Prius, Volkswagen Passat, etc. Authors explain that this is particularly appealing when there are large sources of open-domain text (including the Web), since named entities are well represented on them and it is straightforward to obtain high-quality sets of instances automatically from such sources, among other reasons.

From Section 2 we can conclude that different parameters are required to formulate the rules that precisely determine animacy in Spanish nouns. So we propose to obtain the class of nouns [+H] identifying the contexts where human noun instances appear. For example, John is an instance of lawyers, of Pumas soccer players, of UNAM's workers, etc. These classes (lawyers, players, workers) correspond to nouns with human mark.

### 3.1 Instances

Considering the work of Lin [9], where contexts are used to infer the meaning of an unknown word and are then employed to obtain similar words as an initial step in learning the definition of a word, we may find the contexts where proper names appear and from these obtain the animate nouns occupying the same contexts. For example, let us consider the following immersed phrases in sentences from Mexican newspaper texts:

1. . . . , *el éxito de* ***Francisco Céspedes*** *es indiscutible y . . .*
   '. . . , the success of Francisco Céspedes is indisputable and . . . '
2. *El debate entre* ***Cuauhtémoc Cárdenas*** *y* ***Alfredo del Mazo*** *será cerrado . . .*
   'The debate between Cuauhtémoc Cárdenas and Alfredo del Mazo will be tough'

3. *Si el propio* ***Labastida*** *reconoce que hay tres equipos económicos que . . .*
   'If Labastida himself admits that there are three economic teams that'
4. *. . . , pues fue expulsado por el árbitro* ***Eduardo Gasso*** *al acumular . . .*
   '. . . , since he was expelled by the referee Eduardo Gasso on having accumulated'
5. *. . . golpe de Estado contra el entonces presidente* ***Carlos Andrés Pérez,*** *. . .*
   '. . . coup d'état against the president of that time Carlos Andrés Pérez'

In these examples, the proper names in bold letters can be replaced by a general class. For example, Francisco Céspedes can be replaced by the *singer* animate noun, Cuauhtémoc Cárdenas and Alfredo del Mazo can both be replaced by *candidate*, Labastida by *economist*, etc. The last two examples correspond to appositions where the two nouns represent *class-instance*.

Since we first require a list of instances (simple personal names in opposition to name entities), we could select them from available lexicons, gazetteers or Web-derived lists of names. However, the proper name collection obtained in this way will be limited by the source used. To acquire a much wider list of proper names, we begin with a small list obtained from the Web, then increase this list using a two-step technique:

1. Extract animate nouns by means of patterns according to the linguistic rules for apposition and DOM that we describe in the following section.
2. The animate nouns obtained are used again in apposition and the obtained verbs in DOM patterns also are used to obtain new person names.

### 3.2 Patterns

We propose to extract simple proper names by means of patterns developed from the linguistic phenomena described in Section 2. From the four types described we decide to use patterns for direct object and noun apposition. Knowledge information in the syntactic and semantic levels of sentence analysis, in addition to full sentence context or even paragraph context, are required to determine nouns [+H] according to the verb and pronouns. Since we are interested in simpler methods to determine nouns [+H], we develop patterns where a narrow context is useful. For example:

| | | | |
|---|---|---|---|
| VERB | "a" | PERSON_NAME | (*Veo a Juan* 'I see John) |
| DET | NOUN | PERSON_NAME | (*el poeta Rafael* 'the poet Rafael') |

We apply these patterns to a text collection compiled from Mexican newspapers that are daily published in the WEB. The texts correspond to diverse sections: economy, politics, culture, sport, etc. from 1998 to 2002. The entire text collection has approximately 60 million words [7].

We wrote a program that applies such patterns ensuring that PERSON_NAME corresponds to a name from a list of Mexican person names with 456 elements obtained from the Web. The list has 191 masculine names (e.g.: *Aldo, Alejandro, Alfonso*), 178 feminine names (e.g.: *Abelina, Adela, Adelaida*) and 87 names of indigenous origin (e.g.: *Acamapichtli, Acatl, Acatzin*).

In the first step the following examples were obtained, among many others:
*al administrador* (Martín Ortega)
*al doctor* (Juan Ramón de la Fuente)

*asesinaron a* (José Francisco)
*contestó a* (Miguel López)

For the second step, the contexts like *al administrador, al doctor, asesinar a, contestar a*, etc., were used as patterns and the following names were extracted: *Alfonse, Alger, Álvaro, Amós*, among others. After applying the two-step technique and a manual revision of the results from newspaper texts we obtained a list of 836 person names that we call PERNAM, based on the contexts of 163 animate nouns.

## 4 Results: Set of Nouns [+H]

Using the instances of nouns [+H] it is possible to obtain their surrounding contexts and from the correct ones to automatically identify the nouns [+H] in texts. In many cases, the noun phrase in which the instance is included will be unambiguous and clearly associated with the semantic category. For example, soldier in a clear noun phrase context will always be a noun [+H]. In these cases, the noun phrase alone will be sufficient for the correct determination. In other cases, the context itself is not highly predictive and it will be ambiguous with regard to the semantic class.

*First Step: Extraction of Person's Context*

Google [4] released a collection of n-grams from Web pages. For a better analysis of animacy it would be necessary to examine the full context of every sentence. Nevertheless, in this work we use as corpora the Spanish 5-grams, considering that they are sufficient to capture the diverse structure of noun phrases and direct complements. Our work is as follows:

1. Extracting 5-grams having instances, that is, including person names verified in the PERNAM list
2. Assigning POS to each word without disambiguation and simple unknown words POS assignment.
3. Discarding those 5-grams with no clear cohesion between groups of words, that is, with conjunctions, punctuations, etc.
4. Sorting according to context similarities

An extract of the overall results is presented in Table 1. The first column shows the contexts with their POS, where: SP means preposition (SPC preposition abbreviation, SPS any other preposition), VM means verb (VMP participle, VMM imperative, VMI indicative), NC means common noun (NCF feminine, NCM masculine), PP3 is personal pronoun 3rd person, TDM means define article, AQ0 means qualifying adjective. The second column shows the possible class, i.e., noun [+H]. Column 3 gives the instance of the noun [+H] and column 4 shows the frequency of the 5-grams. All examples shown correspond to the apposition case, but other cases were found with direct object and pronoun *le*.

*Second Step: Validation of Contexts*

Our work is as follows:

**Table 1.** Some results obtained from the Spanish 5-grams

| **Patterns** | | | **NUM** |
|---|---|---|---|
| **NP or VP context** | **Class** | **Instance** | **EX** |
| entrevista al 'interview to'<br>VM[PMI]/NCF SPC | Dr.<br>Ing. 'Eng'r.'<br>escritor 'writer'<br>historiador 'historian'<br>poeta 'poet' | Alejandro Pisanty<br>Felipe Zipitría<br>Miguel Angel<br>Luis Suárez<br>Daniel Bellón | 155<br>71<br>43<br>102<br>181 |
| en los 'in the'<br>SPS PP3/TDM/NCM | premios 'prizes' | Martín Fierro | 54 |
| | premios 'prizes' | Oscar | 595 |
| el notable 'The remarkable'<br>TDM NCM/AQ0 | poeta 'poet' | Luis Francisco | 899 |

1. Each result was verified to make sure that the noun phrase context existed. For example, the following incorrect ones were discarded:
   <S> ahora como *Lola* , lit. 'now such as a Lola,'
   alquiler *apartamentos Santa Cruz de* lit. 'apartment rent Santa Cruz of'
   article thumbnail *Jesús no mira* lit. 'article thumbnail Jesus does not watch'
   mencionada *Norma Andina dispone :* lit. 'mentioned Norm Andean arranges:'
2. Each result consisting of a noun phrase was verified by means of concordance. For example, the following incorrect ones were discarded:
   *de los reyes Alfonso IX* 'of the kings Alfonso IX'
   the kings: masculine plural, Alfonso: masculine singular
   *de malezas Carlos Gomez 03* 'of undergrowths Carlos Gomez 03'
   malezas: feminine plural, Carlos: masculine singular
3. Each result consisting of a verb was verified as having a noun phrase before or after the verb.

After this validation, contexts were used to find possible nouns [+H] and, as the authors in [5] indicated, the most productive pattern was that of noun apposition.

*Third Step: Validation of Nouns*[+H]

The derived class may include a lot of noise. For example: *premios* 'prizes' in column 2 of Table 2 corresponds to a noun [-H], and thus filters or estimation methods are required for knowledge discovery. We propose a very simple method: searching in the Web for the opposite linguistic pattern with a common instance:

1. Noun phrase context:
   For noun validation, a Web search for the verb pattern (VERB "a" NOUN [+H]) was launched. For example, the following incorrect ones were discarded:
   *de metro Pedro de Valdivia* launches"vio a metro" with 1 hit
   *lado íntimo Cecilia Bolocco Angelina* launches "vio a lado íntimo" with 0 hits
2. Verb context
   For noun validation in a verb context, a Web search for the apposition pattern (DET NOUN[+H] *Juan*) was launched. For example:
   *anunció el ingeniero Miguel* launches "el ingeniero Juan" with 1,330,000 hits
   *llega cedido David Lizoain* . launches "el cedido Juan" with 3 hits
   The threshold for accepting noun [+H] was set at 50 snippets.

**Table 2.** An extract of the list of Nouns[+H]

| Noun[+H] | CONTEXT | Instance | # |
|---|---|---|---|
| basquetbolista | , el basquetbolista Tony Parker | Tony | 57 |
| baterista | , el baterista Daniel Torrent | Daniel | 98 |
| batería | , el batería David Dowle | David | 50 |
| Beato | , el beato Juan XXIII | Juan | 55 |
| Bielorruso | , el bielorruso Alexander Hleb | Alexander | 43 |
| Bioquímico | , el bioquímico Héctor Molina | Héctor | 133 |
| Boliviano | , el boliviano Enrique García | Enrique | 65 |
| Boricua | , el boricua Carlos Delgado | Carlos | 56 |
| Brasileño | , el brasileño Carlos Bernardez | Carlos | 41 |
| Brigade | , el brigada Luis Conde | Luis | 75 |

*List of Nouns* [+H]

In Table 2 some portions of the resulting list of nouns [+H] are presented. We can observe that some nouns neglected in traditional dictionaries are present in our results, for example: *basquetbolista, clavadista, perredista*, etc. The total results obtained are 57,808 noun [+H] contexts.

We made a small manual evaluation. We collected three Mexican newspaper articles corresponding to 22/12/04. The texts contain 1,154 words and 74 nouns [+H]. After assigning POS to each word without disambiguation and simple unknown words POS assignment we applied our described results: the list of nouns [+H] and contexts. We obtained 0.77 precision and 0.81 recall, where:

Precision: # of correct noun [+H] detected / # of noun [+H] detected

Recall: # of correct noun [+H] detected / # of noun [+H] manually labeled

We found that among the four nouns bad recognized two cases correspond to nouns appearing in the nouns [+H] list but within a non person context. For example: *con el grado de capitán de Ejército* 'with the degree of captain of Army'. The other two cases correspond to bad proper name detection. For all the non-detected nouns [+H] the context does not help and the nouns do not appear in the list.

## 5 Conclusions

The above linguistic descriptions have shown that animacy for Spanish as in other languages depends on diverse features. The chosen techniques based on morphosyntactic features of animacy have proven to extract the human class well. As we have seen, the specification in nouns and the direct object provide stable structures for animacy even in narrow contexts such as those of the 5-grams. Two experiments have been described above which indicate that instances can be used to capture generalizations which pertain to nouns [+H].

We must emphasize that very frequent nouns [+H] are usually well described in other lexical resources but many others are not described in detail or even neglected. The contributions of this paper are: the attempt to discover animacy noun knowledge from very narrow contexts for Spanish nouns and detecting verbs with an animate subject or animate direct object; they are based on unsupervised methods.

## References


1. Aissen, J.: Differential Object Marking: Iconicity vs. Economy. Natural Language and Linguistic Theory 21(3), 435–483 (2003)
2. Altmann, L.J.P., Kemper, S.: Effects of Age, Animacy, and Activation Order on Sentence Production. Language and Cognitive Processes 21(1), 322–354 (2006)
3. Berenguer, C.R., Cruz Pastor Ferrán, M.: ¿Cuánto dura/tarda la clase de Español?: una reflexión sobre determinados usos verbales en Español. In: Lengua y cultura en la enseñanza del Español a extranjeros. Actas del VII Congreso de ASELE, pp. 397i–406i. Ediciones de la Universidad de Castilla la Mancha (1998)
4. Brants, T., Franz, A.: Web 1T 5-gram Version 1 Linguistic Data Consortium (2006)
5. Fleischman, M., Echihabi, A., Hovy, E.: Offline Strategies for Online Question Answering: Answering Questions before They are Asked. In: Proceedings of the ACL Conference, pp. 1–7 (2003)
6. Foundalis, H.E.: Evolution of Gender in Indo-European Languages. In: Proceedings of the 24th Annual Conference of the Cognitive Science Society, Fairfax, VA, pp. 304–309 (2002)
7. Galicia-Haro, S.N.: Using Electronic Texts for an Annotated Corpus Building. In: 4th Mexican International Conference on Computer Science, ENC, Mexico, pp. 26–33 (2003)
8. Heng, J., Lin, D.: Gender and Animacy Knowledge Discovery from Web-Scale *N*-Grams for Unsupervised Person Mention Detection. In: Proceedings of PACLIC (2009)
9. Lin, D.: Automatic Retrieval and Clustering of Similar Words. In: Proceedings of the 17th International Conference on Computational Linguistics, pp. 768–774 (1998)
10. Orăsan, C., Evans, R.: Learning to Identify Animate References. In: Proceedings of the Workshop on Computational Natural Language Learning, ACL (2001)
11. Orăsan, C., Evans, R.: NP Animacy Resolution for Anaphora Resolution. Journal of Artificial Intelligence Research 29, 79–103 (2007)
12. Øvrelid, L.: Empirical Evaluations of Animacy Annotation. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 630–638 (2009)
13. Paşca, M., Van Durme, B.: What You Seek Is What You Get: Extraction of Class Attributes from Query Logs. In: Proceedings of the International Joint Conference on Artificial Intelligence 2007, pp. 2832–2837 (2007)
14. von Heusinger, K., Kaiser, G.A.: Differential Object Marking and the Lexical Semantics of Verbs in Spanish. In: Kaiser, G.A., Leonetti, M. (eds.) Proceedings of the Workshop Definiteness, Specificity and Animacy in Ibero-Romance Languages, pp. 85–110 (2007)
15. von Heusinger, K., Kaiser, G.A.: The Interaction of Animacy, Definiteness and Specificity in Spanish. In: von Heusinger, K., Kaiser, G.A. (eds.) Proceedings of the Workshop: Semantic and Syntactic Aspects of Specificity, Romance Languages, pp. 41–65. Universität Konstanz, Konstanz (2003)
16. Yamamoto, M.: Animacy and Reference: A Cognitive Approach to Corpus Linguistics. Studies in Language Companion Series, vol. 46. John Benjamins, Amsterdam (1999)
17. Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M.C., Wasow, T.: Animacy Encoding in English: Why and How. In: Proceedings of the 2004 ACL Workshop on Discourse Annotation, pp. 118–125 (2004)