**Development of the Module for Spanish**

Rodolfo Bonino

**1. Introduction**

This work outlines the procedures used by some team members of the Infosur Group of the Universidad Nacional de Rosario (Argentina) to implant the grammars and dictionaries of Spanish into NooJ.

This research group was created by Ph.D. Zulema Solanas in 2004 after the seminaries given by Ph.D. Gabriel Bès, Director of the Groupe de Recherche dans les Industries de la Langue (GRIL) of the Universidad Blaise-Pascal, Clermont-Fd, of France in the Universidad Nacional de Rosario. From then on and following the GRIL methodology, members of our university started to develop this discipline through research projects which led to several doctoral theses (concluded or under development).

As most of our team members' academic backgrounds is aligned with Chomsky's linguistic theories, our research in computational linguistics not only involved a change of focus in the theoretical approach, but also an insight and exploration of software tools, a field of which we had little knowledge then.

In our investigations, we used at first Smorph (Segmentation et Morphologie), a tool created by Salah Aït-Mokhtar and in constant development since 1995 and the post-Smorph module (MPS), created in 1999 by Faiza Abbaci for syntax processing. Then, we included the open access demo Xerox Finite State (XFST), with a limited capacity for data input and output.

As these tools have certain restrictions –Smorph and MPS are test programs which are not on sale nor do they have a user license (our access to them was kindly made available by Dr. Bès and we cannot divulgate that, thus) and the license for the use of the professional version of XSFT is too expensive for our budget and requires an infrastructure which our university does not have– we needed to seek alternative softwares. Most open access freewares available on the Internet work like black boxes showing the data inputs and outpust but do not allow users to manipulate the processes in-between. In contrast, NooJ offers the possibility to create dictionaries and grammars, an advantage which enables us to develop different research projects. Consequently, the Director of the Group assigned me the task to study how to use it.

After studying and using this program for more than a year, I could conclude that NooJ is a tool of great importance to us and a very convenient one to include in our projects due to its aptitude for the processing of natural languages plus the advantages it offers as a freeware tool as well as the possibility to exchange information with other research groups. For this reason, we have organized a study group and a series of studies aimed at doctoral candidates.

In the development process for the module for Spanish, the doctoral candidates Bárbara Méndez and Carolina Tramallino adapted from Smorph the inflectional grammars of nouns and adjectives created by the group[[1]](#footnote-1) under the scope of studies directed by Zulema Solanas; Ph.D. Celina Beltrán designed and implemented the process of automatic transformation of the dictionaries which include these categories, the dictionary of adverbs is an adapted version made by the same author of the Smorph dictionary of Ph.D. Andrea Rodrigo; and the outline of verb dictionary and grammar, which require a whole different processing with regard to that of Smorph, was assigned to me.

The remaining categories are just studied superficially only to fulfill a complete analysis of Roberto Arlt's "La muchacha del atado”, the text model through which we developed the elemental syntactic and productive grammars showing the way the program works were developed.

This work does not provide any reference about how NooJ works. For those who are not familiar with this tool, we recommend studying the tool’s tutorial in Spanish. Reading the manual (in English) and analyzing the modules for other languages will allow us all to evaluate the validity of our proposal to improve it or open other fields of investigation.

**2. Properties**

See the file *\_properties.def* (clic in *open > Properties' definition* [*NooJ > sp > Lexical Analysis*])

As explained before in the tutorial, this file shows the features which are used to perform the lexical tagging. If the tag is applied to all the inflectional variants of a word, it's convenient to associate this tag to the dictionary entry; if the tag corresponds to a single variant, this tag should be associated to that variant in the grammar.

For the analysis of “La muchacha del atado”, in addition to the categorial and morphological features associated in the dictionaries and in the inflectional grammars with the corresponding string of characters, there are used:

a) The feature *apócope* for adjectives such as *buen* as *tan*. In this case, the word is included as a dictionary entry.

b) The dialectal features *RIOP* (as belonging to the Río de la Plata area) and *LUNF* (for *lunfardo*)*.* These features are used to indicate that these words do not belong to regular Spanish.

The first feature is used to indicate standard Río de la Plata words which are usual. In grammar, *verbos.nof*, is applied to the verbal forms used with the Spanish second singular personal pronoun *vos* (known as *voseo*) (*vos cantás* in place of *tú cantas*) and, in the dictionary, for lexical entries which do not correspond to regular Spanish (*yugar* instead of *trabajar (work))*, *pibe* instead of *muchacho, chico (guy*)).

The second applies to forms even marginal to the Río de la Plata variant of Spanish. In the model text, this tag is assigned only to the noun *pebete*, which is a derived form of *pibe* but its use is more restricted. As this follows the derivational pattern of *vejete*, it can analysed through a derivational grammar, but we prefer to include it directly into the dictionary because the form appearing in the text is, at the same time, a derived one (pebetito).

The case of the *voseo* does not give rise to any instances of doubts with regard to the tag RIOP because this has displaced the Spanish second personal pronoun form known as *tuteo* and is the only form which is used at present. The other cases present a greater degree of controversy, as *pibe* is not used in formal register and *yugar* is not usual nowadays.

**3. Inflectional grammars**

The models used by the dictionaries to analyze and generate all the category forms which have morphological variations are created in inflectional grammars. In Spanish, we can state two major models: the nominal model and the verb model which is the more complex and applies only to verbs. The first model can be common for nouns and adjectives, but as each category is the object of study of two doctoral theses, two separate grammars are created.

Adverbs, conjunctions, prepositions and clitics are invariable forms in Spanish. Determiners, quantifiers, pronouns, relative forms, interrogative forms and exclamative forms have invariable forms, variable forms following the nominal grammar (in which case it is used the grammar of the nouns) and special variable forms; in this case, each form has an entry in the dictionary.

**3.1.** **Verb inflection**

See the file \_*verbos.nof* (clic in *open > Grammar* [*NooJ > sp > Lexical Analysis*])

In the issue number 5 of *Revista Infosur* (online) I presented a proposal for the implantation of verb morphology into NooJ which, with some variants, was applied to the analysis of "La muchacha del atado". The following is an overview of the methodology used:

In order to define an identity in NooJ, it is possible to use primary elements (with no previous definition) as previous definitions. The primary elements of verb morphology could be verb bases and the endings, each of them represented by a specific string of characters. As the morphological tags belong to the items, in *amará*, -*ará* should take the tag for the indicative *future 3rd person singular* and that same tag should be applied to the ending -*erá* in *temerá*. That is to say, each tag should be applied to different strings of characters according to the verb conjugation of Spanish.

However, from the observation of the system, it is possible to define a set of characters as previous definitions, which are more adequate and economic for the processing. As a result we can distinguish the following verb formants:

**a. endings:** these are elements which are common to all the regular verbs and a high number of irregular verbs. Endings are used as previous definitions with the corresponding tag. For example, *TER1 = á* with the tag *futuro, indicativo, 3era persona, singular* is applied to all the verbs with regular future formation (*amar* - *amará*)as well as those verbs with irregular future formation (*tener – tendrá*).

**b. intermediate segments:** these present variations according to the conjugation but they are recurrent in the system.

**c. bases:** they are the maximum amount of infinitive characters which can be bundled with the intermediate segments and the endings.

We avoid the use of such concepts as root, thematic vowel and desinence as these belong to the field of non-computational linguistics and these theoretical notions do not coincide with the operational concepts used here: The infinitive base, the future tenses and the Spanish simple conditional tense is, in general, the complete infinitive bundled with an empty element *<E>* or with different endings; when the thematic vowel is the same as that of the infinitive, it is part of the base so that in order to obtain it, only one character is removed from the infinitive. In other cases, the base is the infinitive with one or two characters less, to which they can be added as intermediate segments thematic or desinence vowels which keep the regularity in the endings. For example, to generate the third person singular for the past perfect simple of the indicative mood, two characters are deleted in the infinitives of the first conjugation and the ending (*am-ó*) is added directly, For the same tense and mood for the second conjugation, two characters are deleted and an *i*-intermediate segment is added and then is followed by the ending (*tem-i-ó*); and for the third conjugation, a character is deleted and the ending is added (*parti-ó*). This strategy is the one which allows us to use one ending in the Spanish tripartite verb conjugation.

The endings are grouped in sets which can be applied to the three conjugations and to a wide range of irregular verbs, too. The task of creating the models becomes complex when we apply the criterion of not using the same tag for different elements but it quantitatively reduces the amount of tagged elements. In this way, each tag for tense, mood, person and number is assigned to the ending; the tags assigned to empty elements *<E>* are used when the base, bundled or not with intermediate segments, matches with a verb form. This form at the same time can be bundled with other endings to generate other verb forms. For example, in *amar*, the base matches with the infinitive form but if the *á-*ending is added, it is obtained the third person singular of the future of the indicative and thus, the tag *inf* is assigned to the empty element *<E>* and the tag *fut+ind+p3+s*, to *a*.

Most of the verb irregularities take place at the base of this word class. In these cases, it will be necessary to indicate to NooJ which are the changes it has to make besides adding the intermediate segments and the endings. When the irregularity takes place at the desinence, it is not possible to add the endings. Therefore, the used method is to run a direct tagging of the string of characters following the base.

**3.2.** **Nominal inflection**

See the files \_*adjetivos.nof* and \_*nombres.nof* (clic in *open > Grammar* [*NooJ > sp > Lexical Analysis*])

As stated above, these grammars are Smorph’s grammars which have been adapted to fit into NooJ. In the *Tutorial*, you can view the general criteria used to create them.

In the case of gender-or-number invariable nouns and adjectives, the criterion used is to tag them as masculine and feminine, and singular and plural respectively.

Replacing these tags by the *invariable* tagin grammar is a very simple task and enough for an automatic modification of the dictionary outputs. If this option is selected, you have to take into account that it is not possible to use the same tag for gender and number; two different tags should then be created, for example: *geninv* (for invariable gender) and *numinv* (for invariable number).

**4. Dictionaries**

See the files \_*adjetivos.dic,* \_*nombres.nof, \_adverbios.dic, \_verbos.dic, \_otrascategorías.dic* (clic in *open > Dictionary* [*NooJ > sp > Lexical Analysis*])

As we consider that any research in computational linguistics requires electronic dictionaries as complete as possible as a point of departure, the goal in the development of the Spanish module is, chiefly, to put at other researchers’ hand the basic work carried out by our group in order to join efforts to work towards the study of our language without starting over again. For this reason, the dictionaries are limited to generate the morphological variants of the lexical entries and occasionally, the dialectal marks (RIOP and LUNF) but they do not include the tags which we use in our individual works to calculate the syntactic projection.

In order to distribute the tasks among the members of the group, there were created dictionaries defined by categories. The dictionary of nouns contains 9298 entries from which NooJ generates 19731 inflected forms, the dictionary of adjectives contains 2725 entries from which NooJ generates 10873 inflected entries, the dictionary of adverbs contains 393 invariable entries and the dictionary of other word categories contains 273 entries from which NooJ generates 318 inflected forms. In “other categories”, there are included 44 determiners (articles, demonstratives and possessive prenonominal forms), 4 subordinating conjunctions, 7 coordinating conjunctions, 18 prepositions and 2 contractions, 11 clitics, 8 relatives and 8 interrogative-exclamative forms.

Although these dictionaries do not pick all the entries of the RAE Dictionary, they far exceed the vocabulary of the text taken here as model. Instead, the dictionary of verbs contains only 280 entries (17907 inflected forms) including the verbs appearing in the text "La muchacha del atado".

The increase and the enhancement of these dictionaries is a collective task and for which we expect to work together with other collaborators to continue carrying it out.

**5. Productive grammars**

See the file \_*npropio.nom*, \_*enclíticos.nom, \_enclítico1.nom* (clic in *open > Grammar* [*NooJ > sp > Lexical Analysis*])

As the model text is a short text, the proper names could be included in the dictionaries; however, this option seems to be not adequate at all because if the corpus is enlarged, the amount of proper names will exceed the possibilities of creating a dictionary. As a consequence, we have created a productive grammar which acknowledges as proper name any sequence of characters beginning with a capital letter. This grammar will only be effective only if a low level of priority is given and the dictionaries are complete; otherwise, any sequence beginning with capital letters will be tagged as a proper name (See *Tutorial*).

The enclitic particles of the infinitives (*decirlo, decirle, decírselo*), the gerunds (*diciéndolo, diciéndole, diciéndoselo*), the imperatives (*dilo, dile, díselo*) and the subjunctive forms used as affirmative imperatives (*dígalo, dígale, dígaselo*) present a serious problem for processing them in NooJ: the synthetic spelling (with no space between the verb and the clitic form) hinders their analysis through a syntactic grammar, and the use of a morphological grammar would imply multiplying the verb models according to the clitic forms which they admit. This would involve, for example, creating a model for *cantar*, which can have an accusative case (*cantarlo*), a dative case (*cantarle*) or both forms (*cantárselo*) and another model for *escapar* which has the same morphology (regular verb with the first conjugation) but can take the quasi-reflexive enclitic (*escaparse*), the dative of origin (*escaparle*) or both forms (*escapársele*) but will rarely take the accusative case (*escaparlo*). In addition to this difficulty, there are others which appear with the orthographic accentuation. A task with such magnitude requires a specific study.

At this level, we create only two productive grammars which might account for this problem with different levels of efficiency:

The first (enclíticos.nom) recognizes any cluster of letters followed by *me, te, se, nos, le, les, lo, la, los, las*, or the possible combinations between two of them as a verb + clitic sequence. Similar to the previous case, this grammar will be efficient only if it is assigned a low priority level and the dictionaries are complete. Otherwise, it can tag a word such as *caramelo* as the verb sequence *cara* followed by the clitic forms *me* y *lo*.

The second (enclíticos1.nom) is more accurate as it indicates that any sequence of letters which is recognized as V (verb) followed by any sequence of letters recognized as CL (clitic) is given two tags: the segment identified as verb is tagged as V with its corresponding lemma and the segment recognized as clitic is given the tag CL. The problem with this grammar lies in the fact that it does not recognizes the sequence *olvidándose* because, when taking the enclitic particle, the gerund takes a written accent. Thus, NooJ finds the character string *olvidándo* as not matching any recongnized verb form. When a word has two enclitic particles, all the accents on the second syllable move to the second to the last syllable of the word with the corresponding change of written accent, this grammar is not suitable to process it.

**6. Syntactic Grammars**

See the files \_*tiempocompuesto.nog*, \_*vozpasiva.nog*, \_*períf\_mod.nog, \_períf\_mod\_conj.nog, \_períf\_mod\_prep.nog, \_períf\_temp\_asp\_a.nog, \_períf-temp\_asp\_de.nog, \_cardinales.nog* and *\_sintagmanominal.nog*

The periphrases of Spanish compound tenses are in the border between the syntactic and the morphological levels, but for the automatic analysis, the only option is to process them at the syntactic level. Consequently, the first syntactic grammar created is that which allows us to see the verbal conjugation of the simple tenses with that of the compound ones.

The goal is to build a grammar which can recognize a syntactic object whose morphological properties (person, number, time and tense) matching the Spanish auxiliary verb *haber* and whose lemma matches that of the infinitive. To achieve this, it is necessary the use of variables.

The grammar of compound time is used as a base to create the grammar of the passive voice so that their simultaneous use can recognize sequences such as *ha leído, ha sido* y *ha sido leído.*

Following the same criterion, it is possible to create grammars for Spanish gereund and infinitive periphrases which can analyse sequences such as *poder leer, empezar a leer*, which at the same time can be the base of other periphrases with a more complex structure such as *puede ser leído* or *empezó a ser leído*. We have developed grammars for modal and temporal periphrasis more frequent in Spanish.

There is another crossroad between morphology and syntax as regards numeral forms. In Spanish, there are simple numeral forms (*veinte*) and complex numeral forms which can be written in one word (*veintitrés*) or in two or more words (*treinta y tres*, *novecientos treinta y tres mil*). For the processing of complex cardinal numbers which are written in two or more words, we have created a grammar which recognizes the numeral forms from *uno* to *novecientos noventa y nueve mil*. A similar grammar could be used for ordinal forms and, despite its jumbled design, could have the capacity to generate an endless array of cardinal and numeral forms.

Additionally, a simple grammar is added, which recognizes nouns (mesa), determiner + agreeing noun sequences (*la mesa, las mesas*) and determiner + adjective sequences (*la verde, el verde, lo verde*) or numerals (*las diez, el diez*) as noun syntagm.

**7. Conclusion**

We do not ignore the limitations of our work, but we have the conviction that the progress of a discipline is a creation which is done in a collective fashion and that NooJ is a powerful tool for the management of natural languages, which can be used in different research projects. For this reason, we provide useful dictionaries and basic grammars for other users interested in the computational study of the Spanish language

**Desarrollo del módulo de español**

Rodolfo Bonino

**1. Introducción**

En este trabajo se detallan los procedimientos utilizados por algunos integrantes del Grupo Infosur de la Universidad Nacional de Rosario (Argentina) para implantar en NooJ gramáticas y diccionarios del español.

El grupo de investigación fue creado por la Dra. Zulema Solana en 2004, a partir de los seminarios que el Dr. Gabriel Bès, director del Groupe de Recherche dans les Industries de la Langue (GRIL) de la Universidad Blaise-Pascal, Clermont-Fd, Francia, dictó la Universidad Nacional de Rosario. Desde ese momento, siguiendo la metodología del GRIL, se comenzó a desarrollar la disciplina en nuestra universidad a través de proyectos de investigación que dieron lugar a varias tesis doctorales (concluidas o en preparación).

La formación previa de la mayoría de los integrantes está centrada en la lingüística chomskiana, por lo que la investigación en lingüística computacional, además de un cambio de perspectiva teórica, implicó explorar el campo de las herramientas informáticas, hasta ese momento vagamente conocido por nosotros.

Inicialmente se utilizó la herramienta Smorph (Segmentation et Morphologie), desarrollada por Salah Aït-Mokhtar desde 1995 y el módulo módulo post-Smorph (MPS), creado en 1999 por Faiza Abbaci para el tratamiento de la sintaxis; posteriormente, se incorporó el demo de libre acceso de Xerox Finite State (XFST) , que tiene una capacidad reducida en lo que respecta al ingreso de datos y a las salidas que muestra.

Dado que estas herramientas tienen ciertas limitaciones –Smorph y MPS son programas de prueba que no están a la venta ni tienen licencia de uso (nuestro acceso a ellos es gentileza del Dr. Bès y, por ello, no podemos divulgarlo) y la licencia para el uso de la versión profesional de XFST tiene un costo que no está al alcance de nuestro presupuesto y requiere de una infraestructura que no tiene nuestra universidad–, se impuso la necesidad de buscar programas alternativos. A diferencia de la mayoría de los programas gratuitos y de libre acceso disponibles en Internet, que funcionan como cajas negras, que muestran las entradas y las salidas, pero no permiten manipular el proceso intermedio; NooJ ofrece la posibilidad de crear diccionarios y gramáticas y, por lo tanto, desarrollar diversos proyectos de investigación. Consecuentemente, la directora del grupo me encomendó que me dedicara a estudiar el modo de utilizarlo.

Al cabo de más de un año de estudio y uso del programa; pude concluir que la aptitud de NooJ para el tratamiento del lenguaje natural, sumada a las ventajas que implican la gratuidad, el libre acceso y la posibilidad de intercambio con otros grupos de investigación, hace que se convierta en una herramienta de gran importancia, que nos conviene incorporar en nuestros proyectos. Por esta razón, hemos organizado grupos de estudio y estudios dirigidos para doctorandos.

Para el desarrollo del módulo del español, las gramáticas flexionales de sustantivos y adjetivos de Smorph, creadas por el grupo[[2]](#footnote-2) durante los años 2005 y 2006, fueron adaptadas a la sintaxis de NooJ por las doctorandas Bárbara Méndez y Carolina Tramallino, en el marco de los estudios dirigidos; la Dra. Celina Beltrán diseñó y aplicó el procedimiento de transformación automática de los diccionarios que incluyen estas categorías; el diccionario de adverbios es una adaptación, realizada por su autora, del diccionario Smorph de la Dra. Andrea Rodrigo; y el planteo de la gramática y el diccionario de verbos, que requieren un tratamiento totalmente diferente al de Smorph, estuvo a mi cargo.

Las categorías restantes reciben un estudio superficial, al solo efecto de lograr el análisis completo de texto "La muchacha del atado" de Roberto Arlt, que se presenta como modelo y, a partir del cual, también se desarrollaron gramáticas productivas y sintácticas elementales que muestran el funcionamiento del programa.

Dado que aquí no se explica el funcionamiento de NooJ, se sugiere a los lectores no familiarizados con la herramienta, estudiar previamente el tutorial en español. La lectura del manual (en inglés) y el análisis de los módulos de otras lenguas, permitirá evaluar la validez de nuestra propuesta y mejorarla o abrir otros campos de investigación.

**2. Propiedades**

Ver el archivo *\_properties.def* (clic en *open > Properties' definition* [*NooJ > sp > Lexical Analysis*])

Como se explicó en el tutorial, en este archivo se declaran los rasgos que se utilizan para etiquetar el léxico. Si la etiqueta se aplica a todas las variantes flexivas de una palabra, conviene asociarla a la entrada del diccionario; si corresponde a una sola variante, se debe asociar a esa variante en la gramática.

Para el análisis de "La muchacha del atado", además de los rasgos categoriales y morfológicos, asociados en los diccionarios y en las gramáticas flexivas con la cadena de caracteres que les corresponde, se utilizan:

a) El rasgo *apócope* para adjetivos como *buen* y cuantificadores como *tan*. En este caso, la palabra se incluye como una entrada del diccionario.

b) Los rasgos dialectales *RIOP* (riplatense) y *LUNF* (lunfardo)*.* Se utilizan para marcar palabras que no corresponden al español general.

El primero se utiliza para palabras usuales en español rioplatense estándar. En la gramática *verbos.nof*, se aplica a las formas verbales del voseo (*vos cantás* en lugar del general *tú cantas*) y, en el diccionario, para entradas léxicas que no corresponden al español general (*yugar* por *trabajar*, *pibe* por *muchacho, chico*).

El segundo se aplica a formas que son marginales incluso en el español rioplatense, en el texto se asigna únicamente al sustantivo *pebete*, que es una forma derivada de *pibe*, pero de uso más restringido. Como sigue el patrón de *vejete*, podría tratarse mediante una gramática derivacional, pero preferimos incluirlo directamente en el diccionario porque la forma que aparece en el texto es, a su vez, derivada (*pebetito*).

El caso del voseo no presenta dudas con respecto a la etiqueta RIOP porque ha desplazado completamente al tuteo y es la única forma que se emplea en la actualidad. Los otros casos son más controversiales, dado *pibe* no se utiliza en el registro formal y *yugar* actualmente es poco usual.

**3. Gramáticas flexionales**

En las gramáticas flexionales se crean los modelos que utilizan los diccionarios para analizar y generar todas las formas de las categorías que tienen variaciones morfológicas. En español, podemos establecer dos grandes modelos: el verbal, que es el más complejo y se aplica solo a los verbos; y el nominal. Este último podría ser común para los sustantivos y los adjetivos; pero, dado que cada una de las categorías es objeto de estudio de dos tesis doctorales diferentes, se crean dos gramáticas separadas.

Los adverbios, las conjunciones, las preposiciones y los clíticos son invariables. Los determinantes, los cuantificadores, los pronombres, los relativos, los interrogativos y los exclamativos presentan formas invariables, formas variables que siguen la gramática nominal (en cuyo caso se utiliza la gramática de los sustantivos) y formas variables especiales; en este caso cada forma tiene una entrada en el diccionario.

**3.1.** **La flexión verbal**

Ver el archivo \_*verbos.nof* (clic en *open > Grammar* [*NooJ > sp > Lexical Analysis*])

En el número el número 5 de *Revista Infosur* (en línea), presenté una propuesta para la implantación de la morfología verbal en NooJ, que, con algunas variantes, es la que se aplicó para el análisis de de "La muchacha del atado". A continuación se reseña la metodología utilizada:

Para definir una entidad en NooJ, se pueden utilizar tanto elementos primarios (que no han sido definidos previamente) como definiciones previas. Los elementos primarios de la morfología verbal podrían ser las raíces y las desinencias verbales, cada una de ellas representadas por una determinada cadena de caracteres; como las etiquetas morfológicas corresponden a la desinencia, en *amará*, *ará* debería llevar la etiqueta *futuro, indicativo, 3era persona, singular* y esa misma etiqueta se debería aplicar a *erá* en *temerá*; es decir, cada etiqueta se aplicaría a distintas cadenas de caracteres, según la conjugación verbal.

Sin embargo, a partir de la observación del sistema, es posible definir conjuntos de caracteres como definiciones previas, que resultan más adecuados y económicos para el tratamiento; por lo tanto, se distinguen los siguientes formantes verbales:

**a. terminaciones:** son los elementos comunes a todos los verbos regulares y a un gran número de verbos irregulares. Se utilizan como definiciones previas con la etiqueta correspondiente. Por ejemplo *TER1 = á* con la etiqueta *futuro, indicativo, 3era persona, singular* se aplica tanto a los verbos con futuro regular (*amar* - *amará*)como a los que tienen futuro irregular (*tener – tendrá*).

**b. segmentos intermedios:** presentan variaciones según la conjugación, pero son recurrente en el sistema.

**c. bases:** son la máxima cantidad de caracteres del infinitivo que se pueden concatenar con los segmentos intermedios y las terminaciones.

Se evita utilizar los conceptos de raíz, vocal temática y desinencia, provenientes de la lingüística no computacional, porque esas nociones teóricas no coinciden con los conceptos operacionales aquí empleados: la base del infinitivo, los futuros y el condicional simples es, generalmente, el infinitivo completo, concatenado con un elemento vacío *<E>* o con diferentes terminaciones; cuando la vocal temática coincide con la del infinitivo, forma parte de la base, de modo que para obtenerla se le borra un solo carácter al infinitivo; en otros casos, la base es el infinitivo menos uno o dos caracteres, a los que se pueden agregar como segmentos intermedios vocales temáticas o desinenciales que permiten conservar la regularidad de la terminaciones; por ejemplo, para generar la tercera persona del singular del pretérito perfecto simple del indicativo, a los infinitivos de la primera conjugación se le borran dos caracteres y se les agrega directamente la terminación (*am-ó*); para el mismo tiempo de la segunda se le borran dos caracteres y se le agrega un segmento intermedio *i* seguido de la terminación (*tem-i-ó*) y para la tercera conjugación se borra un carácter y se agrega la terminación (*parti-ó*); esta estrategia es la que permite utilizar una sola terminación en las tres conjugaciones.

Las terminaciones se agrupan en conjuntos que sean aplicables tanto a las tres conjugaciones regulares como a un amplio grupo de verbos irregulares. El criterio de no aplicar la misma etiqueta a distintos elementos complejiza la tarea de crear los modelos, pero reduce cuantitativamente la cantidad de elementos etiquetados. De este modo, cada etiqueta de tiempo, modo, persona y número se asigna a la terminación; las etiquetas asignadas a elementos vacíos *<E>* se utilizan cuando la base, concatenada o no con segmentos intermedios, coincide con una forma verbal; esta forma, a su vez, puede concatenarse con otras terminaciones para producir otras formas verbales; por ejemplo, en *amar*, la base coincide con la forma de infinitivo, pero, si se agrega la terminación *–á*, se obtiene la tercera persona del singular del futuro del indicativo, por lo tanto, la etiqueta *inf* se asigna al elemento vacío *<E>* y la etiqueta *fut+ind+p3+s*, a *a*.

La mayoría de las irregularidades verbales se dan en la base, en estos casos será necesario indicarle a NooJ cuáles son las modificaciones que debe hacer, además de agregar los segmentos intermedios y las terminaciones. Cuando la irregularidad se da en la desinencia, no es posible agregar las terminaciones, por lo tanto, se recurre a etiquetar directamente la cadena de caracteres que sigue a la base.

**3.2.** **La flexión nominal**

Ver los archivos \_*adjetivos.nof* y \_*nombres.nof* (clic en *open > Grammar* [*NooJ > sp > Lexical Analysis*])

Como se explicó más arriba, estas gramáticas son una adaptación a NooJ de las gramáticas Smorph. En el *Tutorial* se pueden ver los criterios generales utilizados para su creación.

En los casos de sustantivos y adjetivos invariables en género o número, para simplificar el tratamiento sintáctico, se ha adoptado el criterio de etiquetarlos como masculinos y femeninos, singulares y plurales, respectivamente.

El reemplazo en la gramática de estas etiquetas por la de *invariable* es muy sencillo y será suficiente para que se modifiquen automáticamente las salidas del diccionario. Si se elige esta opción, se deberá tener en cuenta que no es posible utilizar la misma etiqueta para el género y el número; por lo tanto de deberán crear dos etiquetas diferentes; por ejemplo, *geninv* (género invariable) *numinv* (número invariable).

**4. Diccionarios**

Ver los archivos \_*adjetivos.dic,*  \_*nombres.nof, \_adverbios.dic, \_verbos.dic, \_otrascategorías.dic* (clic en *open > Dictionary* [*NooJ > sp > Lexical Analysis*])

Dado que consideramos que cualquier investigación en lingüística computacional requiere, como punto de partida, de diccionarios electrónicos lo más completos posible; el objetivos del desarrollo del módulo del español es, fundamentalmente, poner disposición de otros investigadores el trabajo básico realizado por nuestro grupo con la finalidad de aunar esfuerzos para avanzar en el estudio de nuestra lengua sin tener que comenzar desde cero. Por esta razón, los diccionarios se limitan a generar las variantes morfológicas de las entradas léxicas y, ocasionalmente, las marcas dialectales (RIOP y LUNF), pero no incluyen las etiquetas que utilizamos en nuestros trabajos individuales para calcular la proyección sintáctica.

Con la finalidad de distribuir las tareas entre los integrantes del grupo, se crearon diccionarios por categorías. El diccionario de sustantivos contiene 9298 entradas, a partir de las cuales NooJ genera 19731 formas flexionadas; el de adjetivos, 2725 entradas, que generan 10873 formas flexionadas; el de adverbios cuenta con 393 entradas invariables y el de otras categorías tiene 273 entradas, que generan 318 formas flexionadas. En "otras categorías se incluyen" 44 determinantes (artículos, demostrativos y posesivos prenominales), 94 cuantificadores (definidos, indefinidos y numerales cardinales que se escriben en una sola palabra), 76 pronombres (personales, demostrativos y posesivos posnominales), 4 conjunciones subordinantes, 7 conjunciones coordinantes, 18 preposiciones y 2 contracciones, 11 clíticos, 8 relativos y 8 interrogativos-exclamativos.

Aunque no recogen todas las entradas del diccionario de la RAE, estos diccionarios exceden ampliamente el léxico del texto que se presenta como modelo; en cambio, el diccionario de verbos contiene solo 280 entradas (17907 formas flexionadas), entre las que se cuentan los verbos que aparecen en el texto "La muchacha del atado**".**

El incremento y la mejora de estos diccionarios es una tarea colectiva para cuya realización esperamos contar con nuevos colaboradores.

**5. Gramáticas productivas**

Ver el archivo \_*npropio.nom*, \_*enclíticos.nom, \_enclítico1.nom* (clic en *open > Grammar* [*NooJ > sp > Lexical Analysis*])

Como se trata de un texto breve, los nombres propios podrían incluirse en los diccionarios; sin embargo, esta opción no parece la más adecuada porque, si se amplía el corpus, la cantidad de nombres propios excedería las posibilidades de confeccionar un diccionario. Consecuentemente, se ha creado una gramática productiva, que reconoce como nombre propio cualquier secuencia de caracteres que comienza con mayúscula. Esta gramática será eficaz solo si se le asigna una prioridad baja y los diccionarios están completos, de lo contrario etiquetará como nombre propio a cualquier secuencia comenzada con mayúscula (Ver *Tutorial*).

Los enclíticos de los infinitivos (*decirlo, decirle, decírselo*), los gerundios (*diciéndolo, diciéndole, diciéndoselo*), los imperativos (*dilo, dile, díselo*) y los subjuntivos empleados como imperativos afirmativos (*dígalo, dígale, dígaselo*) presentan un serio problema para el tratamiento en NooJ: la grafía sintética (sin espacio entre el verbo y el clítico) impide analizarlos mediante una gramática sintáctica y el uso de una gramática morfológica implicaría multiplicar los modelos de verbos según los clíticos que admiten; por ejemplo, habría que crear un modelo para *cantar*, que puede tener acusativo (*cantarlo*), dativo (*cantarle*)o ambos (*cantárselo*) y otro para *escapar* que tiene la misma morfología (verbo regular de la primera conjugación), pero puede llevar enclítico cuasi reflejo (*escaparse*) o dativo de origen (*escaparle*), o ambos (*escapársele*), pero raramente llevaría acusativo (*escaparlo*), a esta dificultad se suman las que traen los cambios en la acentuación ortográfica. Una tarea de tal magnitud requiere un estudio específico.

En esta instancia, nos limitamos a crear dos gramáticas productivas que pueden dar cuenta de este problema con distinto grado de eficacia:

La primera (enclíticos.nom) reconoce cualquier conjunto de letras seguidas por *me, te, se, nos, le, les, lo, la, los, las*, o las combinaciones posibles entre dos de ellos como secuencia de verbo más clítico. Al igual que en el caso anterior, esta gramática será eficaz solo si se le asigna un nivel de prioridad bajo y los diccionarios son completos, de lo contrario etiquetará una palabra como *caramelo* como un secuencia de verbo *cara* seguido por los clíticos *me* y *lo.*

La segunda (enclíticos1.nom) es mucho más precisa porque indica que cualquier secuencia de letras que sea reconocida como V (verbo) seguida de cualquier secuencia de letras reconocida como CL (clítico) recibe dos etiquetas: el segmento identificado como verbo se etiqueta como V con el lema correspondiente y el segmento reconocido como clítico recibe la etiqueta CL.El problema de esta gramática es que no reconoce la secuencia *olvidándose* porque, como, al incorporar el enclítico, el gerundio recibe una tilde; NooJ encuentra la cadena de caracteres *olvidándo,* que no coincide con ninguna forma verbal reconocida. Dado que, cuando tienen dos enclíticos, todas las formas verbales de acentuación grave se transforman en esdrújulas, con el consiguiente cambio de acentuación gráfica, esta gramática resulta inadecuada para su tratamiento.

**6. Gramáticas sintácticas**

Ver los archivos \_*tiempocompuesto.nog*, \_*vozpasiva.nog*, \_*períf\_mod.nog, \_períf\_mod\_conj.nog, \_períf\_mod\_prep.nog, \_períf\_temp\_asp\_a.nog, \_períf-temp\_asp\_de.nog, \_cardinales.nog* and *\_sintagmanominal.nog*

Las perífrasis de tiempo compuesto están en la frontera entre lo sintáctico y lo morfológico; pero para el análisis automático,la única opción es darles un tratamiento sintáctico. Consecuentemente, la primera gramática sintáctica que se crea es la que permite completar la conjugación verbal de los tiempos simples con la de los tiempos compuestos.

El objetivo es construir una gramática que reconozca un objeto sintáctico cuyas propiedades morfológicas (persona, número, tiempo y aspecto) coincida con el auxiliar *haber* y cuyo lema coincida con el del infinitivo; para lograrlo es necesario utilizar variables.

La gramática de tiempo compuesto se utiliza como base para crear la de voz pasiva, de modo que la aplicación simultánea de ambas reconoce secuencias como *ha leído, ha sido* y *ha sido leído.*

Con el mismo criterio, es posible crear gramáticas para perífrasis de infinitivo y gerundio, que analicen secuencias como *poder leer, empezar a leer*, que, a su vez, pueden ser la base para otras más complejas como *puede ser leído* o *empezó a ser leído*. Hasta el momento, hemos desarrollado gramáticas de las perífrasis modales (PM) y temporales-aspectuales (PTA) más frecuentes.

Otro punto de intersección entre la morfología y la sintaxis es el de los numerales: en español existen numerales simples (*veinte*) y complejos, que se pueden escribir en una sola palabra (*veintitrés*) o en dos o más palabras (*treinta y tres*, *novecientos treinta y tres mil*). Para el tratamiento de los números cardinales complejos que se escriben con dos o más palabras se ha creado una gramática que reconoce los numerales de *uno* a *novecientos noventa y nueve mil*. Una gramática similar se podría utilizar para los ordinales y, salvando la dificultad de lo abigarrado del diseño, tendría capacidad para generar una serie infinita de cardinales y ordinales.

Adicionalmente, se agregan una gramática simple, que reconoce sustantivos (*mesa*), secuencias de determinante y sustantivo concordados (*la mesa, las mesas*) y secuencias de determinantes y adjetivos (*la verde, el verde, lo verde*) o numerales (*las diez, el diez*) como sintagma nominal.

**7. Conclusión**

No ignoramos las limitaciones de nuestro trabajo; pero tenemos la convicción de que el progreso de una disciplina es una creación colectiva y de que NooJ es una herramienta poderosa para el tratamiento de las lenguas naturales, que puede ser utilizada en diversos proyectos de investigación. Por ello, ponemos al alcance de otros usuarios interesados en el estudio computacional de la lengua española diccionarios y gramáticas básicas que puedan servir de base para su labor.

1. BOLLA, Liliana; BONINO, Rodolfo; FREIDENBERG, María Susana; GASPARINI, Rosana; KOCKAK, Claudia; KOZA, Walter; MORO, Stella Maris; RIVERO, Silvia; RODRIGO, Andrea; VALENTI, Viviana. Directed by: SOLANA, Zulema. Scientific assessment: BÈS, Gabriel. [↑](#footnote-ref-1)
2. BOLLA, Liliana; BONINO, Rodolfo; FREIDENBERG, María Susana; GASPARINI, Rosana; KOCKAK, Claudia; KOZA, Walter; MORO, Stella Maris; RIVERO, Silvia; RODRIGO, Andrea; VALENTI, Viviana. Dirección: SOLANA, Zulema. Asesoramiento científico: BÈS, Gabriel. [↑](#footnote-ref-2)