AnaPro, tool for identification and resolution of direct anaphora in Spanish

I. Toledo-Gómez[1], E. Valtierra-Romero[1], A. Guzmán-Arenas[2], A. Cuevas-Rasgado[1], L. Méndez-Segundo[1]

[1] Escuela Superior de Cómputo ESCOM-IPN
[2] Centro de Investigación en Computación CIC-IPN
México, D. F., México

ABSTRACT

AnaPro is software that solves direct anaphora in Spanish, specifically pronouns: it finds the noun or group of words to which the pronoun refers. It locates in the previous sentences the referent or antecedent which the pronoun replaces. An example of a direct anaphora solved is: the pronoun "he" in the sentence "He is sad."

AnaPro directly supports text analysis (to understand what a document says), a non trivial task since there are different writing styles, references, idiomatic expressions, etc. The problem grows if the analyzer is a computer, because they lack "common sense" (which persons possess). Thus, before text analysis, its preprocessing is required, in order to assign tags (noun, verb,…) to each word, find the stems, disambiguate nouns, verbs, prepositions, identify colloquial expressions, identify and resolve anaphora, among other chores.

AnaPro works for Spanish sentences. It is a novel procedure, since it is automatic (no user intervenes during the resolution) and it does not need dictionaries. It employs a heuristic procedure to discover the semantics and help in the decision. Its results are good (81-100% of correct answers); nevertheless, more tests will give a better idea of its goodness.

Keywords: I.2. Artificial Intelligence, I.2.7 Natural Language processing, Text Analysis, Anaphora resolution.

RESUMEN

AnaPro es un software que resuelve problemas con anáforas directas en español, especialmente pronombres: la herramienta encuentra el sustantivo o grupo de palabras al cual se refiere el pronombre. Localiza en oraciones previas la referencia o antecedente y lo reemplaza por el pronombre que le corresponde. Un ejemplo de anáfora directa resuelta es el pronombre "El", en la oración "El está triste".

AnaPro soporta directamente análisis de textos (para entender lo que el documento dice), esta es una tarea no trivial ya que existen diferentes formas y estilos de escribir, referenciar, expresiones regionales (mexicanismos, argentinismos, etc). Todavía se complica más si el analizador es una máquina, porque no tiene "sentido común" (que cualquier persona puede tener). Así, antes de analizar el texto, se requiere pre-procesarlo para asignar etiquetas (sustantivos, verbos, …) a cada palabra, encontrar las derivaciones, desambiguar sustantivos, verbos, preposiciones, identificar expresiones coloquiales, identificar y resolver anáforas entre otras tareas.

AnaPro trabaja para oraciones en español. Es un proceso novedoso porque es totalmente automático (no requiere de la intervención de un usuario) y tampoco necesita de diccionarios. Emplea un procedimiento heurístico para descubrir la semántica y apoyarse en las decisiones que debe de tomar. Ha resultado ser bueno (con un 81-100% de aciertos); sin embargo, se siguen introduciendo más ejemplos nuevos que le darán más fortaleza en sus respuestas.

1. Introduction

The Spanish Royal Academy defines anaphora as a "kind of deixis (the specific function of some word, as *this, those, I, there, above, yesterday, now*) that some words perform (for instance, *herself* in *Sally preferred the company of herself*) to pick up the meaning of a part of discourse (or previous text) already expressed." [Diccionario, Edición XXII, Online source 7]. To resolve an anaphora is to know to which part of the discourse the deixis is pointing. Webster's dictionary is clearer: "It is the use of a grammatical substitute (as a pronoun or a pro-verb) to refer to the denotation of a preceding word or group of words; *also*: the relation between a grammatical substitute and its antecedent."

Anaphora resolution is very important in Natural Language Processing (NLP). In every case text

coherence is paramount. So it is in Project OM* (Ontology Merging), which uses frames [6] to represent the context of the theme (of a Spanish document), with the goal to transform that text into an ontology that comprises its contents. This paper describes AnaPro, one of the modules of OM*. AnaPro determines the direct anaphora (when the deixis are pronouns). Its output eases the building of the part of the ontology that corresponds to the text being analyzed. For instance.

El perro entró dentro del ***bote****.* ***Él*** *se deslizó tristemente.*

{The dog entered the **boat**. **It** sadly slided away}.

A person easily finds that the sliding was done by the boat, but a program may have difficulty to find out to which of the nouns perro {dog} or bote {boat}, the pronoun **él** {he, it} refers to. In the example, a person asks herself: *who can slide away?* The boat, since the dog can not. This is so because a person has common sense, experience, previous knowledge, etc. For the computer, analysis and verification of several information sources and tools (frames, dictionaries, thesaurus, etc.) is usually needed in order to have a satisfactory solution.

Anaphora resolution has been an important task for NLP during years. Because of its importance, the problem haws been tackled from several fronts. For instance, [8] use semantic criteria, as described in §3.

Morales [7] and others have proposed methods that try to resolve anaphora resolution (both direct and indirect) in Spanish, using dictionaries. Their methods do not solve all grammatical forms.

Nevertheless, simple methods using limited knowledge can be applied in many cases with satisfactory results. For instance, text from a restricted domain or culled from a controlled vocabulary. An example is Martínez [4], who solves anaphora with the help of the SUPP (Slot Unification Partial Parser) analyzer, which provides lexical and syntactic information to find out the antecedent of higher relevance in the list of antecedents (one of then is the referent). Although with some limitations, the work proposes interesting improvements to the algorithm, as described in §3.

Another approach suggests to use linguistic and structural knowledge which are domain-independent, as in [9]. It deals with anaphora in Spanish identified by third person pronouns and adjectival anaphora. Described in §3, it uses a statistical model to determine the complements of verbs, adjectives and some nouns, based on a corpus.

An important work, since it deals with unrestricted domains, is that of [2]. It locates and extracts information where answers to questions can be inferred. It does this in Spanish and in English. It is briefly described in §3.

AnaPro only solves direct anaphora, particularly pronoun resolution. Of the already mentioned previous works, only Sidorov in [9] resembles our solution.

The paper is thus organized: §1 describes the state of the problem and related previous work; §2 gives the theoretical framework; §3 presents work related to resolution of direct anaphora and places AnaPro in this context. §4 details our solution and solves some examples. §5 explains how AnaPro works; §6 provides more examples taken from “real life” (Online sources).

2. Theoretical framework

Anaphora resolution has been the focus of attention of philosophers, linguists, cognitive and artificial intelligence scientists, psycholinguists and computational linguists. This is important because the anaphora:

- It is one of the most complex phenomena in natural language.
- It has been shown that syntactic, semantic and pragmatic factors take part in it.

Its resolution is needed in a wide range of NLP tasks: natural language interfaces, document understanding [for the computer to answer complex questions about the knowledge that lies in them (this is the goal of the aforementioned Project OM*)], machine translation, information extraction, automatic summarization, finding differences or contradictions between statements on the same subject in two documents ....

Two important types of anaphora exist: direct and indirect anaphora.

Direct anaphora associates a pronoun or reference to some entity already mentioned in the text. Fort instance, *los alumnos estaban contentos pues ellos habían pasado a la semifinal* {students were happy because they had passed to the semifinal}; here the pronoun *ellos* {they} refers to *los alumnos* {the students}.

Indirect anaphora establishes an associative link between a linguistic entity (word or phrase) and an entity previously introduced through the text or the speech [7]. For example, *Eliseo entró a un circo, el payaso era muy simático* {Eliseo went to a circus, the clown was very nice}. In this example the discourse focuses on the circus and it is the previous entity or antecedent to which *payaso* {clown} belongs or refers. A connection exists between *payaso* and the circus where Eliseo went, which has at least one clown. This information comes from the extension of the context by encyclopedic knowledge (common sense); as a consequence, the whole sentence renders a consistent interpretation, as implied by the relevance principle, used above, or by the scenario approach (frames, [6]), where the use of *circo* {circus} invokes a scenario (the frame of circus) that implicitly contains at least one clown. AnaPro does not analyze indirect anaphora.

2.1 Contributions

AnaPro converts a Spanish text into another text without direct anaphoric references. The anaphors are solved, replacing for instance *ellos* {they} by *estudiantes* {students}. The final text, without anaphoric references, is an input to be used in project OM. Some of its contributions are:

1) The focus of AnaPro is to find (automatically) the noun that gives coherence to a text. Other works such as Morales, [7] require a user. Instead, Martinez, 2002, needs an information detection system to determine the most relevant preceding topics.

2) Another important point of AnaPro is that it resolves anaphora in complete texts, not in short dialogues (Martinez, 2002). Thus, full text can be entered into the tool for its analysis and resolution.

For anaphora resolution and replacement of the pronoun by its corresponding part of discourse, we must understand what we mean when by the coherence of a text.

2.2 Text coherence

The word *text*, from Latin *textus*, means weaving, interlacing something. What distinguishes a text from a sequence of loose sentences is the fact that in the text, information interleaves and intertwines, producing a unitary or unified sense, as a logical thread. This, which is called semantic coherence, guides us when we determine the theme or context that the text describes, and through which sentences lose ambiguity. Text coherence is text consistency.

Text coherence is the property that points to the information to be communicated and how it is to be done, in what order, with what degree of accuracy or detail, with which structure. Broadly speaking, we can say that coherence is the property responsible for the quantity, quality and structuring of information. It is basically semantic; therefore, it affects the true meaning of the text. To resolve direct anaphora, AnaPro uses the following linguistic model.

2.3 Linguistic model

The model used to solve direct anaphora generally contains these four steps:

1. Description of the references, starting with the determiners that function as common markers of the different relationships in the speech;
2. Identification of the nominal ellipsis (removal of the noun kernel, see § 2.4) which affects the function of the determiners since they are forced to perform as if they were extrinsic pronouns;
3. Recording the referential expression and entity concepts as support to explain the reference and its variants: direct and indirect correference, and direct and indirect anaphora; and
4. Resolution of the reference.

Since the former points mention determiners, it is necessary to define them.

2.4 Determiners

Determiner is a term that denotes the lexical unit that precedes a noun in a nominal phrase to specify its reference, including the quantity of the noun. For instance,

1.- Todos esos vehículos están en venta. {All those vehicles are for sale}

The determiner is ***Todos*** {all} which determines the quantity, and ***esos*** {those} is a demonstrative pronoun that indicates the place, relative to the emitter, where the vehicles are.

In general, the determiners give rise to the definite expressions where these parts of speech are used: the definite articles (*el*, *la*, *lo*, *las*, *los*) {the, it}; the demonstratives (*aquel*, *aquella*, *aquellas*, *aquellos*, *esa*, *esas*, *ese*, *esos*, *esta*, *estas*, *este*, *estos*, *tal*, *tales*, *semejante*, *semejantes*) {that, that, those, those, that, those, that, those, this, these, this, these, such, such, like, like}; the possessive determiners (*cuya*, *cuyas*, *cuyo*, *cuyos*, *mi*, *mis*, *nuestra*, *nuestras*, *nuestro*, *nuestros*, *su*, *sus*, *vuestra*, *vuestras*, *vuestro*, *vuestros*, *tu*, *tus*, etc.) {whose, whose, whose, whose, my, my, our, our, our, our, his or her, his or her, your, your, your, your, your, your, etc.}, and the quantifiers (*todo*, *algún*, *cada*) {all, any, each}. With any other determiner, an expression is considered an indefinite expression (see examples 11 and 15).

Let us talk now about the nominal ellipsis, the second step in the linguistic model.

2.5 The nominal ellipsis

If we define syntagma as a group of words, then a noun syntagma or noun phrase is a group of words (in a sentence) that plays the role of a noun. Since our problem is anaphora resolution, this noun phrase is also called *core.*

Noun ellipsis (omission) is within the limits of the noun phrase. In it, the core is not expressed (does not appear), and the syntagma is represented by the remaining modifiers. In the following example, X marks the suppressed element (alumnos) {students} and the place of the ellipsis:

2. Hoy vienen los alumnos de tercero; mañana los X de segundo. {Today, the third grade students come; tomorrow, the second grade X}

Generally, the noun phrase occurs in the ellipsis; nevertheless, there are cases such as: *Tengo dos relojes digitales de cuarzo que me han traído*. *Te regalo uno*. {I have two digital quartz watches that somebody brought me. I give you one}.

Some times, the selection of elements taken from the preceding syntagma by the missing syntagma depend on extralinguistic factors. For instance, in the answer (to the above offer): "*No, gracias; ya tengo yo otro*". {No, thank you, I already have another}. In [3], point to a sequence of precedences when it is necessary to take some non core element from the antecedent. The sequence (from higher to lower probability) is: restrictive modifier (prepositional syntagma), adjective, and quantifier. Every noun ellipsis contains some new information that is precisely the difference from its antecedent. Sometimes the phonetic emphasis can guide the rejection of an element from the antecedent noun ellipsis:

3. *"Te fumaste 20 cigarrillos." –"Me fumé 10."* {You smoked 20 cigarettes. --I smoked ten}.

The noun ellipsis is a linguistic economy mechanism, because it can be inferred or understood in the context of the discourse; no linguistic entity appears that must be linked with an antecedent; simply, an empty place is left, signaling within brackets the omitted information; for instance:

4. *Juan dibujó una casa y Pedro [dibujó] una oveja.* {John drew a house and Peter [drew] a sheep}.

5. *Juan toca el piano; María [toca] la guitarra*. {John plays piano; Mary [plays] guitar}.

Ellipsis takes its full name according to the missing element. Thus, in former examples 4 and 5 it is known as verb ellipsis because the verb is the missing element. In this paper we focus on noun ellipsis, because it alters the normal use of the determiner, as the following examples show:

6. *Juana compró una lavadora nueva, pero María compró una [lavadora] de segunda mano*. {Jane bought a new washing machine, but Mary bought a second hand [washing machine]}.

7. *El compositor favorito de Juan es Bach, pero el [compositor favorito] de José es Handel*. {The favorite composer of John is Bach, but that [favorite composer] of Joseph is Handel}.

The determinants, *una* {a} and *el* {the} in examples 6 and 7, fulfill the determining function upon a noun or a noun phrase, in the first example, but the ellipsis allows their omission in example 7. In this case the determiner functions as if it were an extrinsic pronoun, affecting the original function of the determiner in the sentence.
The third point of the linguistic model refers to the referential expression; let us talk about it now.

2.6 The referential expression

Considering the entity as the concept associated with a linguistic reference, let us paraphrase Saussure "the association between the signified and the signifier." The referential expression can be defined as the linguistic structure (expression) allowed to the emitter (or author) in order to introduce, or to mention again, the entities in the discourse.

The discourse is given in the communication act, through the text, when the emitter (or writer) introduces and discusses the entities (individuals, objects, events, actions, states, relations or attributes), be they abstract or concrete. Although several types of referential expressions do exist, such as the locative and temporal *pro formas*, AnaPro is limited to the nominal referential expressions; that is to say, to the resolution of direct anaphora through pronouns.

An important part of the algorithm given here is the detection and marking of the referential expressions, since they depend on the functioning of the lexical units of the determiner type, which can represent different functions within the sentence, in addition to being affected by the phenomena of ellipsis. As an initial example, we present the determiners *la* {the} and *una* {a}.

8. *Juan$_1$ baña a la niña$_2$ y José$_3$* ***la$_2$*** *seca con la toalla$_4$*. {John$_1$ bathes the girl$_2$ and Joseph$_3$ dries **her**$_2$ with the towel$_4$}.

In this example, the first occurrence of *la* fulfils the function of the determiner (definite article), but in the second occurrence, it functions as an accusative pronoun, singular, third person.

9. *Juan$_1$ compró una paleta$_2$ de dulce. María$_3$ también compró* ***una$_4$***. {John$_1$ bought a lollipop$_2$. Mary also bought **one**$_4$}.

In example 9, the first occurrence of *una* fulfils the determiner (indefinite article), but its second occurrence functions as the extrinsic pronoun due to the ellipsis. Let us note that the expression *una*$_4$ refers to the same concept of lollipop but to a different object in the real world. Now we give some examples where the determiners are identified inside parenthesis:

10. Ni él mismo sabía a ciencia *cierta* lo que pasaba. (adjetivo) {He himself did not know *for sure* what was happening} (adjective)
11. Se expresaba con *cierta* dificultad al hablar... (det. Indef.) {He spoked with *certain* difficulty when speaking} (Indefinite determiner)
12. Caminé por las calles solitarias. En *algunas* había faroles... (pron. Indef.) {I walked the empty streets. *Some* had lanterns} (Indefinite pronoun).
13. El techo tiene *algunas* manchas de humedad. (det. Indef.) {The ceiling has *some* water stains} (Indefinite determiner)
14. Juan le entregó una carta a Luis y *otra* a Sofía. (pron. Indef.) {John gave a letter to Luis and *another* to Sofia} (Indefinite pronoun).
15. Lo alcanzaremos en la *otra* calle. (det. Indef.) {We'll catch him in the *other* street} (Indefinite determiner)
16. ¿Corremos a la casa? Si, a la *una*, a las dos y a las tres. (sustantivo) {Do we run home? Yes, *at once*} (Noun).

It is important to see in example 16 the use of the determiner **la** that converts another determiner into noun, in the expression "a la una" {at once}, because in Spanish "everything that can be preceded by a determiner is made noun." Now we present some examples about this phenomenon, showing within parenthesis the original category of the word (in italics) that is being used as a noun, and underlining the determiner.

17. Juan le regaló en su cumpleaños *una* pulsera. (Det. cardinal) {John

gave her *a* bracelet in her birthday}. (Cardinal determiner).

18. Es suficiente acudir de cada tres veces, *una*. (Pron. indef.) {It is enough to go, of each three times, *once*} (Indefinite pronoun).
19. Juan le guardaba desde entonces una gran fidelidad. (Det. indef.) {Since then, John showed *a* great loyalty to her}.
20. Todos los colores me gustan pero el *rojo* es mi favorito. (Adjetivo) {I like all colors, but *red* is my favorite}.
21. Este "*el*" es un artículo y no un pronombre... (Det. def.) {This "*he*" is an article and not a pronoun} (Definite determiner).
22. Tu *reír* me fascina. (verbo inf.) {I love your *laugh*} (Indefinite verb).
23. El *ayer* ya no existe. (adverbio) {*Yesterday* does not exist} (Adverb).
24. Sobra esa *de* en tu oración de la tarea. (Det. def) {That "*d*" is extra in your homework sentence} (Definite determiner).
25. ¿Porqué pone tanto *pero* a este trabajo? (Conjunción) {Why so much *objection* to this work?} (Conjunction).

According to the above, during the reading process (left to right), when a determiner is found, there are three possible cases:

1) There may be a noun, immediately or with additional modifiers interposed, which is the most "normal" case.

2) There may be a word that is fulfilling the function of a noun (it is being used as a noun), and which "normally" belongs to a different category.

3) There may not be a noun, because the determiner is fulfilling another function in the sentence.

These considerations are taken into account in the algorithm that detects and marks algorithm of nominal referential expressions. Special mention deserve the contracted prepositions *al* (*a el*) {to the} and *del* (*de el*) {of the}, which function as determiners and thus should be considered by the mentioned algorithm.

Once the referential expressions are defined, it is time to define the coreference.

2.6.1 The coreference

Coreference is the relation between two noun phrases that are interpreted as referring to the same extralinguistic entity. In linguistic representations, coreference is conventionally denoted by coindexing, for instance:

26. María$_1$ dijo que ella$_1$ vendría. {Mary$_1$ said that she$_1$ would come}

where coreference is denoted by subscript 1 (*ella*) {she} that refers to Mary.

2.7 Reference resolution process

The task of whoever receives the information is to conduct the reference resolution process, which is achieved with the following sequence of steps:

i. To identify as accurately as possible the entity to which the emitter (or writer) is referring to (the referred) with the referential expression.
ii. To determine if this entity has already been previously mentioned (or referred) in the context of the discourse, or if it is new.
iii. If it has already been mentioned (coreference), it will establish the link between referential expressions in the linguistic context; the first entity already mentioned already has the link to the entity in the discourse context.
iv. If it is new, it should be integrated it as part of the linguistic context, creating the link from the referential expression to the entity in the linguistic context.

The whole process must be based on the morphological and syntactic features of the text, which involve the use of determiners, and where the concepts of referential expression, coreference and direct and indirect anaphora (already defined in §2) are closely related.

3. Related work

There are some studies that have addressed the problem of identifying anaphoric references in the text. The focus of those studies, their resolution

algorithms and their limitations are briefly mentioned.

In [8] approach the anaphora in Spanish and resolve it through a sequence of semantic criteria, generating an abductive theory (starting from facts, an hypothesis is generated) for the pronominal anaphora. They propose the following resolution procedure:

1. Obtain the grammatical form of the sentence and determine the main verb.
2. Incorporate into the context the non anaphoric referents that show up in the analysis.
3. Determine the anaphoric terms and their features.
4. Determine, as a function of the type of argument that the anaphora occupies with respect to the main verb, its thematic role.
5. Look for possible referents for the anaphora, among those occurring in the same and former sentences. Each one must pass three filters:
    a. Grammatical agreement (it is the syntactic consistency criterion).
    b. Agreement between the thematic roles (preferential criterion).
    c. Semantic coherence of the resulting sentence (a mixture of preferential and semantic consistency criteria).

Finally the paper presents a solution, although it requires more experimentation to test the hypotheses presented.

Morales [7] focuses on the anaphoric references, both direct and indirect. His resolution algorithm is based in the use of a thesaurus, and in the detection and contextualization of coreferences, in such a way that not all the grammatical forms are processed by the algorithm, but through an entity exclusion process, irrelevant information is thrown away.

The algorithm discovers the existing interrelation between the coreference phenomena (direct and indirect) and the indirect anaphora, and that both use nominal referential expressions to manifest their presence.

It determines the evaluation order required to discriminate the phenomena that are used as basis to detect the presence of the indirect anaphora.

It develops a method based in the scenario model of the linguistic context that models the human reader, needed for the resolution of the indirect anaphora and the coreferences. But one of its shortcomings is that the additional information needed is not automatically obtained. When the system was tuned to use it with free text, its precision was 51%, while the human expert obtained a global average of 60%. With the linguistic context based in the scenario model, it achieves the automatic resolution of the nominal indirect anaphora, with a strong dependence on the additional information added.

Martínez [4] uses as input the slot structure generated by the SUPP analyzer. This structure stores lexical, syntactic, morphologic and semantic information of each element of the grammar, and it also includes an anaphora detection mechanism, as well as another mechanism for its resolution. This last method uses the information from the topic detection system in order to determine the most relevant antecedent among the general topics list. The topics information joins the information obtained from other sources, to get the best antecedent. In the same manner, once the anaphora has been resolved, the antecedent occupies the place left by the expression. This system works on dialogues.

Following is a simple example in which Martínez' algorithm is applied to the resolution of anaphora. It is a small dialogue where the system of topics defines three spaces for anaphoric accessibility: a) the space of the actual intervention; b) the space of the previous intervention; and c) the relevant topic in the current dialogue space. The first and second spaces are directly determined thanks to the labeling of turns characteristic of dialogues. The third space is identified by two coefficients that compute the use or lack of use of an entity; that is to say, the importance or irrelevance of an entity occurring in the current intervention.

The algorithm uses three lists to store the entities that can be an antecedent (according to the previously defined spaces). The *list of current local entities,* Fa, contains the entities occurring in the current intervention. The *list of previous local entities,* Fa', contains the entities that occurred in the last intervention. The *list of general topics,* F, contains the entities relevant in the dialogue, evaluated according to frequency and distribution.

T01: <H1> Buenos días. {Good morning}

T02: <H2> Buenos días. ¿Qué deseas?. {Good morning. What do you want?}
T03: <H1> Quiero **manzanas**. {I want **apples**}
T04: <H2> ¿De qué clase **las** quieres?. {Of what kind do you want **them**?}
T05: <H1> No importa si son buenas. {It does not matter as long as they are good}
T06: <H2> **Éstas las** he recibido **esta mañana**. Son muy buenas. {I have received **these this morning.** They are very good}

The algorithm starts by considering the turns T01 and T02 as continuation turns since they do not contain any entity. Thus, the three lists remain empty until T03. When processing T03, *manzanas* {apples} is considered an entity, and thus introduced in the list of current local entities Fa:
Fa=[manzanas] {apples}
T03 is thus considered an intervention. When it finishes, the entities in Fa are incorporated into the list of general topics F with their corresponding weight, and they are also stored in the list of previous local entities Fa'. After these operations, the lists are:
Fa=[ ]
Fa'=[manzanas] {apples}
F=[(manzanas,10)] {apples, 10}

In T04, after the incorporation of *clase* {kind} as an occurring entity, the first anaphora is detected, as the pronoun *las* {them}. At this point the lists are:
Fa=[clase] {kind}
Fa'=[manzanas] {apples}
F=[(manzanas, 10)] {apples, 10}

To resolve the anaphora, the algorithm searches, first, an antecedent in the list of current local entities (Fa). It finds the entity *clase* {kind}; however, the intervention of the system of restrictions rejects it, since it does not meet the required morphological characteristics (it is looking for a plural antecedent). In this way, since Fa does not contain more candidates, the algorithm will look into the previous local entities list, Fa', finding *manzanas* {apples}. After restrictions are applied, *manzanas* is proposed as antecedent for the pronoun and it is included as an entity in Fa:
Fa=[clase, manzanas] {kind, apples}

After processing the first sentence of T04, Fa' is emptied:
Fa'=[ ]
And when T04 is finished, the lists are:
Fa=[ ]
Fa'=[clase, manzanas] {kind, apples}
F=[(manzanas, 20), (clase, 10)] {(apples, 20), (kind, 10)}
Since T05 is a continuation turn, the lists do not change. After T06, the weight of *manzanas* increases due to new appearances of this instance (introduced by the pronouns), while the entity *clase* {kind} loses weight due to the opposite:
Fa=[ ]
Fa'=[manzanas, esta mañana] {apples, this morning}
F=[(manzanas, 40),(esta mañana, 10), (clase, 9)] {(apples, 40), (this morning, 10), (kind, 9)}

We can see that the anaphora is resolved returning *manzanas* as a valid result, according to the coefficient that indicates the frequency of occurrence of entity *manzanas.*

A shortcoming is that it does not incorporate semantic information. Thus, it can not rule out antecedent candidates, and it proposes the use of Spanish Wordnet [Online source 6] as lexical tool, to learn semantic patterns between subject and verb, and between verb and object. It does not solve any other kind of references in the dialogue.

The relevant feature of **AnaPro** (our work), when compared with the work of Martinez (2002), is that **AnaPro** resolves anaphora in texts and not in short dialogues.

Sidorov [9] propose the development of a resolution system for anaphora generated by third person pronouns and adjective anaphors in the dialogues (this work is the closest to our algorithm). The system is based in a syntactic analyzer that uses an extended grammar, context independent, with unification elements. This program incorporates the results of the research to obtain handling patterns for verbs, adjectives and nouns from Spanish. These results allow quantitative classification of the variations generated by the analyzer, using the weights assigned to the variations, following the values of the subcategorization combinations. The weights of the subcategorization combinations are the result of a syntactic analysis process and of an extraction following a statistical model that allows determining the complements of verbs, adjectives

and some nouns from Spanish, starting from a text corpus.

The Open Domain Question Answering Systems ([2]) are tools capable of getting concrete answers directly from natural language documents over unrestricted domains, in response to precise information needs from users. At least, they try to find and extract those areas (of the documents) from whose contents an answer to a specific question can be obtained (by hand). Thus, those systems try to reduce the time invested by users to find concrete information. They have three modules. The first module is a standard Information Retrieval system, collecting documents that may be relevant to the query. The second module, which uses SUPAR (Slot Unification Parser for Anaphora Resolution[1]), has as input the queries and the retrieved documents. The third module is a Question Answering system, which also interacts with SUPAR. It divides the sentences of the documents in order to find those where the correct answer appears. It is in this second module where the pronominal anaphora is resolved, using a lexical analyzer, a syntactic analyzer and a module that solves anaphors, extra-position, and ellipsis, when the result is a slot structure (refer to the description of Martinez' work at the beginning of Section 3) corresponding to the input sentence.

AnaPro does not use slot structures. Its resolution mechanism is detailed in next section.

As we can appreciate, anaphora resolution is a problem not yet fully solved, thus we have proposed and implemented a solution based in pattern identification and determiners enumeration, which we now show.

4. Related work

AnaPro works in five stages.

Stage 1*:* It receives a text tagged with EAGLE tags [10]. The tagged text is stored in a structure in main memory, and the nominal expressions are identified.

Stage 2*:* The pronominal syntagmas are identified, and the cases of nominal ellipsis are stored.

Stage 3*:* The direct anaphoric references are identified.

Stage 4*:* The anaphoric references go through a filter of grammatical restrictions. Here, candidates (for resolution) are identified. If ties exist, criteria of preferences in the assignment apply, so that only one candidate is chosen for each anaphoric reference.

Stage 5*:* In the text, all pronouns solved in stage 4 are replaced by the chosen candidates.

4.1 Stage 1

A portion of text "La batalla de Puebla" {Puebla's battle} [Online source 9] shows AnaPro in action.

"La batalla de Puebla.

Aunque se considera en todo el país, ésta se celebra especialmente en Puebla, México; y en los estados del Sur de Estados Unidos" {Puebla's battle. Although considered in all the country, it is celebrated especially in Puebla, Mexico, and in the southern states of USA}.

This stage reads the tagged text, storing each word and their tags in separate structures, which are numbered. Also, the nominal expressions are obtained, through the use of the module "To obtain the nominal expression" of AnaPro, which is explained below. See Table 1.

Tagged and numbered text		
AD0FS00 = la {the} NCFS0001 = batalla {battle} AC0MSP2 = de {of} NCFS0003 = puebla {Puebla} Fp4 = . CS5 = aunque {although} P00000006 = se {it is} VMIP3S07 = considera {considered} SPS008 = en {in} DI0MS09 = todo {all}	AD0MS010 = el {the} NCMS00011 = país {country} Fc12 = , PD0FS00013 = ésta {it, this} P030000014 = se {is} VMSI3S015 = celebra {celebrate} RG16 = especialmente {especially} SPS0017 = en {in} NCFS00018 = Puebla {Puebla} Fc19 = ,	NCMS00020 = México Fx21 = ; CC22 = y {and} SPS0023 = en {in} AD0MP024 = los {the} NCMP00025 = estados {states} SPCMS26 = del {of the} NCMS00027 = sur {South} SPS0028 = de {of} NCMP00029 = Estados {States} AC0MPP30 = Unidos {United}

Table 1 showing the tagged text, numbered and distributed in three columns.

Table 2 shows the nominal expressions generated from the tagged and numbered text.

[1] SUPAR also has three modules: lexical analysis, syntactic analysis, and linguistic problem resolution.

Nominal expressions generated
NCFS0001 = la batalla de puebla {Puebla's battle}
NCMS00011 = todo el país {all the country}
NCFS00018 = Puebla
NCMS00020 = México
NCMP00025 = los estados del sur {the southern states}
NCMP00029 = Estados Unidos {United States}

Table 2. Nominal expressions identified in the text. The first characters of the tag mean: N (noun), C(common), F (female), S(singular), M(masculine), P(plural). The table shows the order in which they occur. The tagging is done by the tagger TaggerFTG [5] in a preprocessing step. The nominal expressions are generated by the tagger, we only use several filters to identify and delimit them correctly. They are shown here.

Module "To obtain the nominal expression"

1. Each tagged word is numbered (by the tagger).
2. Using the tags, pronouns are identified, as well as their determiners and auxiliary types, through the comparison between each tag and regular expressions, be they a pronoun, noun, adjective, etc.

For instance, to look for pronouns in the text, which will be tagged as pronominal anaphoric references, the expression shown in Figure1 is used. Figure 1 is expanded to Figure 2, showing that a pronoun must match with one of these expressions. It is required that a pronoun must match with the first two of one of the expressions of Figure 2. The remaining eight symbols in the regular expression will define with further detail each type of pronoun, such as its gender and number, which is also considered when searching for candidates.

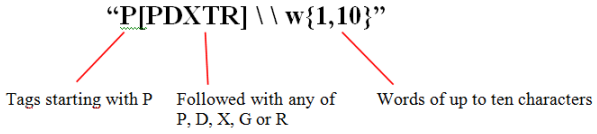


Figure 1. Regular expression used to find pronouns in a text. It is divided in three parts: the start of the tags, the following letters, and the length of the words.

PP********
PD********
PX********
PT********
PR********

Figure 2. Tag types, where asterisks indicate optional areas, and the first two characters indicate the pattern to look for. The standard requires that the tag of a pronoun should have at least eight elements, then we assign an index, which could have from 1 to 10 digits. In this step, the algorithm does not take into account the last 6 or 8 digits, since they provide no information to our tool.

This type of expressions has been used to identify pronouns, articles, nouns, adjectives and in some cases, to distinguish among each particular pronoun, to ascertain its gender and number in order to find pairs of pronouns and nouns.

Once the pronouns or anaphoric pronouns are identified, AnaPro locates the previous portion of text where the possible nouns or determiners may exist, those to which the pronoun refers to. This area or range is known as *nominal expression*, which contains the anaphora and its possible determiners. The name *nominal expression* was changed to *sentence* inside the implementation of AnaPro, for practical purposes.

Function to generate nominal expressions

This part of AnaPro, instead of dealing directly with nouns, starts by correcting labeling errors "on the fly", such as the exchange of a noun label by a quality adjective, in the case of a surname or a name. At the same time, the first step always is to find the nucleus. The process starts going through the text from start to finish in a single cycle. When the first noun nucleus is found, it stores the noun and its tag, then goes back looking for (jumping over phrases or sentences, moving backwards from the found noun) a determiner article, qualifying adjective, number and other modifiers. These are sequentially added as they are found (using a window of at most five elements, which in our tests has never been exceeded). The cycle repeats going forward from the noun, but with a slightly modified order and nouns are sought. An important step in this part is to keep absolutely all the tags of the elements found going forward, in a structure called ACpre. It is used to verify the limits of the previous nominal expressions.

Let us assume that a nominal expression has two qualifying adjectives, and immediately after them there is a "new noun nucleus", that is, a different nominal expression. When the first nominal expression is built, both qualifying adjectives are added, but when the second nominal expression is built, AnaPro first verifies if

the qualifying adjectives have been (or have not) previously assigned. If that is the case, the search backwards for nominal elements is immediately interrupted. The same happens with specific punctuation marks.

An important addition has been the use of date tags as nuclei to construct nominal expressions with historical dates or references to some year, with all that this implies: use of prepositions, numbers, month and years.

Also added is the search for lists and plural expressions, looking for elements such as commas, nouns and conjunctions, in order to generate a single nominal expression, with the tag of the last noun modified to be plural. The search of plurals with conjunctions, unlike all the others, is not very reliable, since errors have been detected when a conjunction is surrounded by two nouns. In this case its effectiveness depends on the correct use of the comma by the writer.

The ACpre structure.

In the algorithm that constructs nominal expressions, the ACpre structure stores trhe names of all the nominal expressions at the same time they are being built, starting from the first word up to the last period. ACpre, a list, is used to avoid the construction of nominal expressions with "name nuclei" that have been previously used as part of a name of another nominal expression. Example: el triunfo del movimiento revolucionario {The success of the revolutionary movement}. A first step renders "triunfo {success} as nucleus, the article "el" {the} is placed before it, and "del movimiento revolucionario" {of the revolutionary movement} is appended as qualifying adjectives. A second step will find "movimiento revolucionario" as nucleus and will try to create a new nominal expression. But, upon analyzing the ACpre structure, it will notice that "movimiento revolucionario" has been used already in a previous expression, thus, it can not be a new nucleus. Thus, it is ignored.

4.2 Stage 2

This stage detects the pronominal syntagmas containing pronouns as nucleus. At the same time all the cases of nominal ellipsis are stored.

Identification of pronominal syntagmas:

PD0FS00013 = ésta {this} (See Table 2)

Here we find only a syntagma as direct anaphoric reference. Other examples of syntagmas are *el cual* {that which}*, la cual* {that which}*, el que* {that what}, which are also identified by AnaPro.

4.3 Stage 3

Candidates are identified, and direct anaphoric references (with gender and number defined) are resolved with a single candidate.

The module of AnaPro that does this task is the ***Classifier.*** It finds the coincidences of pronouns and nouns, classifying them as single or multiple matches (in the case of ties). Finally, it looks for the correct pairs.

Search for candidates.

The tag found in Stage 2 is analyzed:

PD0FS00013 = ésta {this}

It is identified as a female singular pronoun, where the following search pattern has been applied: when searching for candidates:

Index: 13 Pattern: N\w{1}FS\w{3,10} Sentence: 2

indicating that it has to look for female singular nouns in the second or previous sentences. Therefore, the candidate found is (refer to tables 1 and 2):

NCFS0001 = la batalla de puebla{Puebla's battle}

So that we have:

PD0FS00013 = NCFS0001

It means:

ésta {this} = la batalla de puebla {Puebla's battle}

4.4 Stage 4

Candidates are identified, and direct anaphoric references (with gender and number defined) are resolved with a single candidate.

It solves special cases of pronouns and candidates ties. This stage also uses the ***Classifier*****.** Our example does not have special cases or ties among candidates. The following example has them:

Juan compró un boleto en la tienda. Éste estaba roto.
{John bought a ticket in the store. It was torn}

This is a tie, since there are several possible candidates: *Juan* {John}, *boleto* {ticket} and *tienda* {store}. *tienda* can not be because *Éste* {this} is a masculine pronoun and *tienda* has female gender.

Solving the tie

To solve the tie, four important criteria are taken into account:

1. Less distance to the noun.
2. If it is determined.
3. Less distance to the main verb.
4. Higher number of occurrences in the text.

In case the first three criteria are not enough to break the tie, the numbers of occurrences are compared (fourth criteria). For this example, Table 3 shows the weights assigned to each candidate.

Juan compró un boleto en la tienda. Éste estaba roto.

Criterion	Juan {John}	Boleto {ticket}
Less distance to the noun	0	1
Is it determined?	0	1
Less distance to the main verb	1	1
Higher number of occurrences in text	0	0
Total	**1**	**3**

Table 3. Comparison using weights among tied candidates.

When comparing weights among candidates, we obtain ***éste*** {this} = ***boleto*** **{ticket}**, thus solving the sentence as:
Juan compró un boleto en la tienda. [El boleto] estaba roto.
{John bought a ticket in the store. [The ticket] was torn}

4.5 Stage 5

All the solved anaphors are replaced, yielding a new text that keeps coherence. Pronouns are replaced by their respective candidates. The module that performs this task is the ***Resolver,*** which replaces anaphora by their respective noun, generating an output file with the anaphors resolved.

Coming back to previous example:
*"**La batalla de puebla**. Aunque se considera en todo el país, **ésta** se celebra especialmente en Puebla, México; y en los estados del sur de Estados Unidos"* {**Puebla's battle**. Although considered in all the country, **it** is celebrated especially in Puebla, Mexico, and in the southern states of USA}

As result of the ***Resolver,*** the output file would contain:
*"**La batalla de puebla**. Aunque se considera en todo el país, **[la batalla]** se celebra especialmente en Puebla, México; y en los estados del sur de Estados Unidos"*
{**Puebla's battle**. Although considered in all the country, [**the battle**] is celebrated especially in Puebla, Mexico, and in the southern states of USA}

Although AnaPro resolves ***ésta*** {this} as referring to ***la batalla de puebla*** {Puebla's battle}, when replacing (and to keep text coherence), only the nucleus of the nominal expression is replaced. The expression ***la batalla de puebla*** has as nucleus ***la batalla*** {the battle}, which is used in the replacing algorithm.

5. Using AnaPro

AnaPro has two versions: a console and a graphical version. Before using any of them, a file in .txt format is needed, in order to generate the tagged file Taller FT, that returns an extension *.txt.tag*. The Java virtual machine must be already installed, version 1.2 or higher. We use the graphical version to show AnaPro in action.
For instance, in the document about the piano [Online source 1]:
El piano está compuesto por una caja de resonancia, a la que se ha agregado un teclado, por donde se percuten las cuerdas de acero con macillos forrados de fieltro, por éstos se clasifica como instrumentos de percusión. Las primeras composiciones específicas para este instrumento surgieron alrededor del año 1732, entre ellas destacan las 12 sonatas para piano de Lodovico

Giustini. {The piano is composed of a sounding board, to which a keyboard has been added, through which the steel chords are struck with hammers lined with felt; for this reason, it is classified as a percussion instrument. The first specific compositions for this instrument arose circa 1732. These include the 12 piano sonatas of Lodovico Giustini}.

With the help of the example, the functions of the graphic interface (Figure 3) are explained.

1. File menu, to open files with extension .txt.tag. They should be located in the same folders than the original text.
2. Tag showing original text, as found in the file.
3. Tag showing anaphoric references found by AnaPro, the word indices inside the text, and the noun corresponding to each reference; in other words, its solution.
4. Tagged text, used as a reference to compare it with the final tagged text.
5. Auxiliary view of tagged text, to compare it with the anaphoric references.
    a. It shows the nominal expressions generated by AnaPro. Noun + adjectives, article.
6. It shows the new tagged text with the anaphors solved.
7. It shows the text with the words already replaced.
8. Erase all windows.
9. It stores the final file into the same folders containing the initial text file.

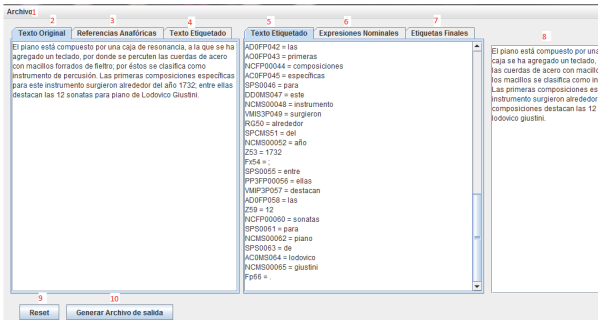


Figure 3 Graphical interface of the resolver.

Figure 4 shows the tag showing the anaphoric references. The three sections marked A, B and C are:

Section A contains the anaphoric references found in the text, indicating the tag and then the word(s).

Section B shows the index of the references, followed by the search pattern (if it is a masculine pronoun, feminine, singular, etc.), and then it shows the sentence where the anaphora was found.

Section C shows the solution to each anaphora found by the resolver. When the ***Generar Archivo*** {generate file} button is clicked, a new file is produced, adding at the end of its name **_ANA.** Thus, if its name is **Mezcal.txt**, the file generated will be **Mezcal_ANA.txt**.

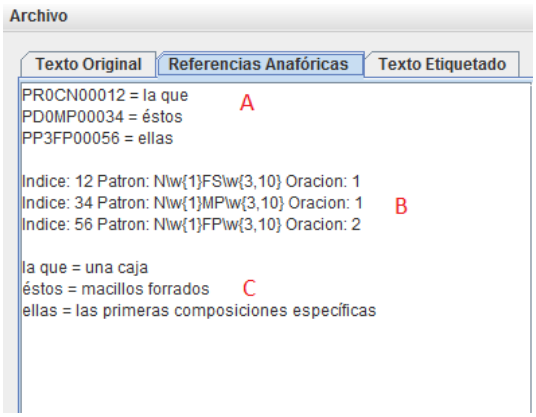


Figure 4 – Tag: Anaphoric references.

6. Tests

Testing has been done using the following texts taken from the Web:

1. "El piano" {the piano} [Online source 1],
2. "El burro flautista" {the flutist donkey} [Online source 2],
3. "Batalla de Puebla" {Puebla's battle} [Online source 3],
4. "Pancho Villa" {Pancho Villa} [Online source 4] y
5. "Mario Almada" {Mario Almada} [Online source 5],

In all of them, AnaPro resolved automatically all the pronominal anaphors found. The same was done manually, in order to verify the effectiveness of the program. The formula is:

$$\text{Percentage of correct responses} = \frac{(\text{number of anaphors correctly solved})}{(\text{total number of anaphors in the document})} \quad (1)$$

Table 4 shows the results obtained using formula 1).

Text	Number of words	Anaphors solved by hand	Anaphors solved by Anapro	% of success
El piano	61	3	3	100
El burro flautista	97	1	1	100
Batalla de puebla	513	7	5	71.4
Pancho Villa	634	11	7	63.6
Mario Almada	601	9	5	55.5

Table 4 where anaphors are solved, both by AnaPro and by hand

Failures, in most cases, were caused by final tagging. Thus, tagging has been manually corrected (to assume a perfect tagger). Retesting gives the results in Table 5.

Text	Number of words	Anaphors solved by hand	Anaphors solved by AnaPro	% of success (with the help of a perfect tagger)
El piano	61	3	3	100
El burro flautista	97	1	1	100
Batalla de Puebla	513	7	7	100
Pancho Villa	634	11	9	81.8
Mario Almada	601	9	8	88.8

Table 5 Texts analyzed number of words, manual solution, automatic solution (assuming a perfect tagger) and Precision of AnaPro.

In this second result (Table 5), failures are due to phrasing errors, or to very elaborate cases or very specific situations in a particular discourse (use of complex cohesion techniques), because the reference becomes incoherent. An example of these cases is the following fragment from Mario Almada [Online source 5]:

"…fácilmente es expuesto a rodajes y en la Ciudad de México empieza a trabajar en un centro nocturno llamado Cabaret Señorial que era propiedad de su padre. Pero antes de esto, cuando su hermano Fernando empezó…" {… easily is exposed to filming and in Mexico City [he] starts to work in a night club called Cabaret Señorial that belonged to his father. But before this, when his brother Ferdinand started…}

The word *esto* {this} does not refer to a noun or a candidate, but to a whole context previously cited; thus, it is difficult to produce a computer tool that can resolve it.

6.1 Analyzing the failures if AnaPro

AnaPro has a high resolution success, except in some cases. Another example is taken from the same text [Online source 5].

"Ha aparecido en más de 200 películas, muchas de las cuales eran drama clásico como su primer film Madre Querida (1935). En ésta apareció de niño con su hermano" {He has appeared in more than 200 films, many of which were classic drama, as his first film Beloved Mother (1935). In this he came out as a boy with his brother.}

Here we have a phrasing error, because the text uses a foreign word (film) that in Spanish is of masculine gender (el filme), but the text refers to it as ésta {this, feminine}. AnaPro generated the nominal expression (a candidate) *"film Madre Querida"* and assigned it the masculine gender. That fools AnaPro, which looks further backwards for a feminine candidate.

AnaPro heavily relies in the results of the tagger, since the anaphoric references fail if the tagger fails. If it is manually corrected (or by using a better tagger), the resolution algorithm works quite well.

Conclusions and future work

A method has been developed, using regular expressions, to separate, enumerate and identify pronominal syntagmas, special cases of nominal ellipsis and candidates to resolve the anaphoric references.

An algorithm has been designed and tested, able to solve ties using only information present in the text, statistical analysis of occurrences, text distance and specific wording criteria, in order to assign weight to multiple candidates.

Former method and algorithm are part of AnaPro, which resolves direct anaphora given by pronouns that occur in Spanish texts. The texts do not refer to a particular topic. AnaPro is part of Project OM* (a follow over to the Ph. D. thesis of one of the authors, [1]), that tries to build the ontology of the knowledge contained in a text, in order to answer complex queries.

In order to minimize errors and to improve the efficiency of AnaPro, it is necessary to perfect the tagger, widening the scope of nouns handled, in order to solve quicly the simpler problems. Also, another tagger that uses EAGLE standard can be used.

Acknowledgements

To ESCOM-IPN, where authors I.T. and E.V. defended their thesis #**20110083**, which gives a more detailed description of AnaPro. Work herein reported was partially sponsored by CONACYT Grant #128163 (Project OM*), by IPN (A.G. as Resident Scientist), and by SNI.

References

[1] A. Cuevas, “Merging of ontologies using semantic properties”, Thesis (Ph. D.) CIC-IPN (On line) México, 2006, In spanish. Available at: htttp://148.204.20.100:8080/bibliodigital/ShowObject.jsp?idobject=34274&idreposiorio=2&tpe=recipiente.

[2] A. Ferrández et al. “¿Cómo influye la resolución de la anáfora pronominal en los sistemas de búsqueda de respuestas?”, RUA. Repositorio Institucional de la Universidad de Alicante. Revistas. Procesamiento del Lenguaje Natural No. 26, ISSN. 1135-5948, septiembre 2000, INV-PLN-Artículos de Revistas, España. Web site http://hdl.handle.net/10045/1907 (Last consulted August 21, 2012) In Spanish.

[3] M. Haliday, and H. Ruqaiya “*Cohesion in English* Longman”, Group United Kingdom, ISBN-10: 0582550416, ISBN-13: 978-0582550414, July 1976, London

[4] P. Martínez “Resolución Computacional de la anáfora en diálogos: Estructura del discurso y conocimiento lingüístico”, RUA. Repositorio Institucional de la Universidad de Alicante. Revistas. Procesamiento del Lenguaje Natural No. 28, ISSN. 1135-5948, mayo 2002 INV-PLN-Artículos de Revistas. España. Web site http://rua.ua.es/dspace/handle/10045/1851 (Last consulted August 21, 2012) In Spanish.

[5] J. Meneses and M. García, "Construction of an analyzer of colloquial sentences and a syntactic tagger of sentences of a document”. B. Sc. thesis No. 2010-0075 ESCOM-IPN. May 2011. In Spanish. (The name of the tagger is TaggerFT).

[6] M. Minsky “A Framework for Representating Knowledge”. MIT-AI Laborarorio Memo 306, Junio 1974 Reprinted in *The Psychology of Computer Vision*, P. Winston, ed. McGrawHill, 1975. Link http://web.media.mit.edu/~minsky/papers/Frames/frames.html (Last consulted August 21, 2012) In Spanish.

[7] R. Morales, “Resolución automática de la anáfora indirecta en español”, Thesis (Ph. D.) CIC-IPN. México. In Spanish, 2004.

[8] F. Salguero and F. Soler “Resolución abductiva de anáforas pronominales”, *IV Jornadas Ibéricas*, Fénix Editors. In Spanish, 2010.

[9] G. Sidorov and O. Olivas, “Resolución de anáfora pronominal para el español usando el método de conocimiento limitado”. Avances en la Ciencia de la computación, 7° congreso internacional ENC-2006, México, In Spanish, 2006, pp. 276–281.

[10] A. Toral et al. “EAGLES compliant tagset for the morphosyntactic tagging of Esperanto”. In *Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing* (RANLP). Borovets, Bulgaria, 2005.

Online source 1
Piano description Archive. Available from: (online): http://es.wikipedia.org/wiki/Piano

Online source 2
Burro flautista description Archive. Available from: (online): http://www.juegosyeducacion.com/fabulas/el_burro_flautista.html

Online source 3
Batalla de Puebla description Archive. Available from: (online): http://www.educar.org/comun/efemerides/Mexico/5demayomexico

Online source 4
Francisco Villa description Archive. Available from: (online): http://www.biografiasyvidas.com/biografia/v/villa.htm

Online source 5
Mario Almada description Archive. Available from: (online): http://es.wikipedia.org/wiki/Mario_Almada

Online source 6
Spanish WordNet 3.0 description Archive. Available from: (online): http://sinai.ujaen.es/timm/wiki/index.php/Spanish_WordNet_3.0

Online source 7
Real Academia Española Dictionary description Archive. Available from: (online): http://buscon.rae.es/drae/?type=3&val=anafora&val_aux=&origen=REDRAE

Annex A. Tables

Tagged and numbered text		
AD0FS00 = la {the} NCFS0001 = batalla {battle} AC0MSP2 = de {of} NCFS0003 = puebla {Puebla} Fp4 = . CS5 = aunque {although} P00000006 = se {it is} VMIP3S07 = considera {considered} SPS008 = en {in} DI0MS09 = todo {all}	AD0MS010 = el {the} NCMS00011 = país {country} Fc12 = , PD0FS00013 = ésta {it, this} P030000014 = se {is} VMSI3S015 = celebra {celebrate} RG16 = especialmente {especially} SPS0017 = en {in} NCFS00018 = Puebla Fc19 = ,	NCMS00020 = México Fx21 = ; CC22 = y {and} SPS0023 = en {in} AD0MP024 = los {the} NCMP00025 = estados {states} SPCMS26 = del {of the} NCMS00027 = sur {South} SPS0028 = de {of} NCMP00029 = Estados {States} AC0MPP30 = Unidos {United}

Table 1 showing the tagged text, numbered and distributed in three columns.

Nominal expressions generated
NCFS0001 = la batalla de puebla {Puebla's battle} NCMS00011 = todo el país {all the country} NCFS00018 = Puebla NCMS00020 = México NCMP00025 = los estados del sur {the southern states} NCMP00029 = Estados Unidos {United States}

Table 2. Nominal expressions identified in the text. The first characters of the tag mean: N (noun), C(common), F (female), S(singular), M(masculine), P(plural). The table shows the order in which they occur. The tagging is done by the tagger TaggerFTG [5] in a preprocessing step. The nominal expressions are generated by the tagger, we only use several filters to identify and delimit them correctly. They are shown here.

Criterion	Juan {John}	Boleto {ticket}
Less distance to the noun	0	1
Is it determined?	0	1
Less distance to the main verb	1	1
Higher number of occurrences in text	0	0
Total	1	3

Table 3. Comparison using weights among tied candidates.

Text	Number of words	Anaphors solved by hand	Anaphors solved by Anapro	% of success
El piano	61	3	3	100
El burro flautista	97	1	1	100
Batalla de puebla	513	7	5	71.4
Pancho Villa	634	11	7	63.6
Mario Almada	601	9	5	55.5

Table 4 where anaphors are solved, both by AnaPro and by hand

Text	Number of words	Anaphors solved by hand	Anaphors solved by AnaPro	% de aciertos (with the help of a perfect tagger)
El piano	61	3	3	100
El burro flautista	97	1	1	100
Batalla de Puebla	513	7	7	100
Pancho Villa	634	11	9	81.8
Mario Almada	601	9	8	88.8

Table 5 Texts analyzed number of words, manual solution, automatic solution (assuming a perfect tagger) and Precision of AnaPro.

Annex B. Figures

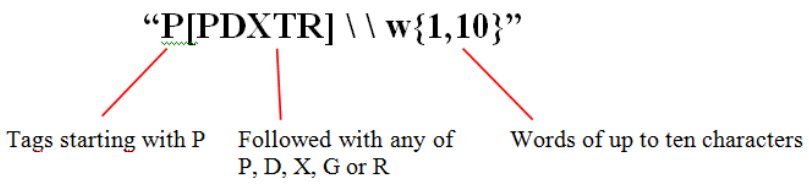


Figure 1. Regular expression used to find pronouns in a text. It is divided in three parts: the start of the tags, the following letters, and the length of the words.

PP********
PD********
PX********
PT********
PR********

Figure 2. Tag types, where asterisks indicate optional areas, and the first two characters indicate the pattern to look for. The standard requires that the tag of a pronoun should have at least eight elements, then we assign an index, which could have from 1 to 10 digits. In this step, the algorithm does not take into account the last 6 or 8 digits, since they provide no information to our tool.

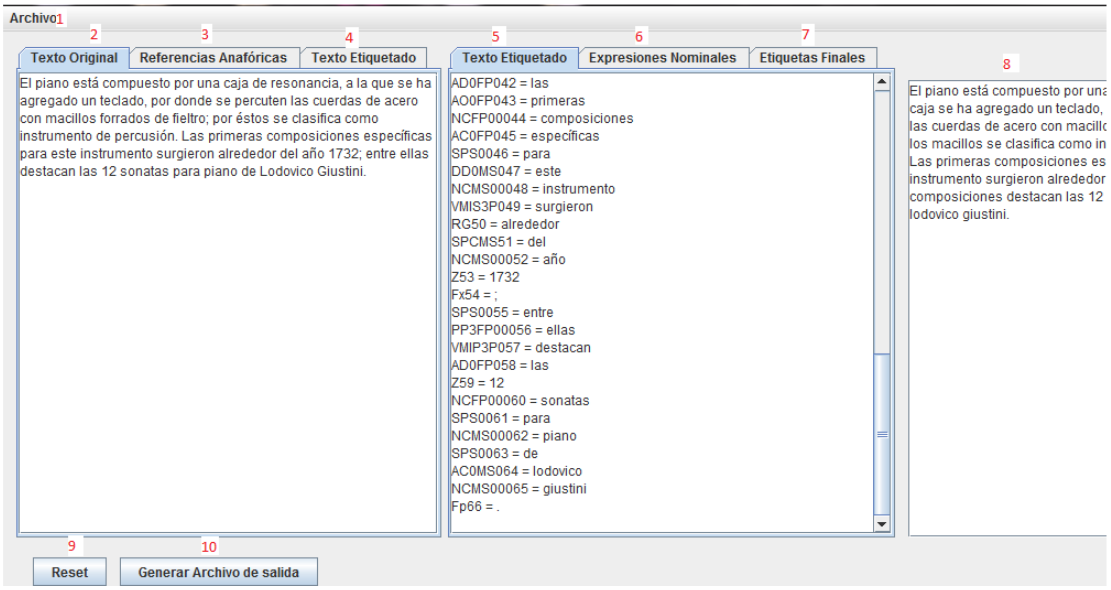


Figure 3 Graphical interface of the resolver

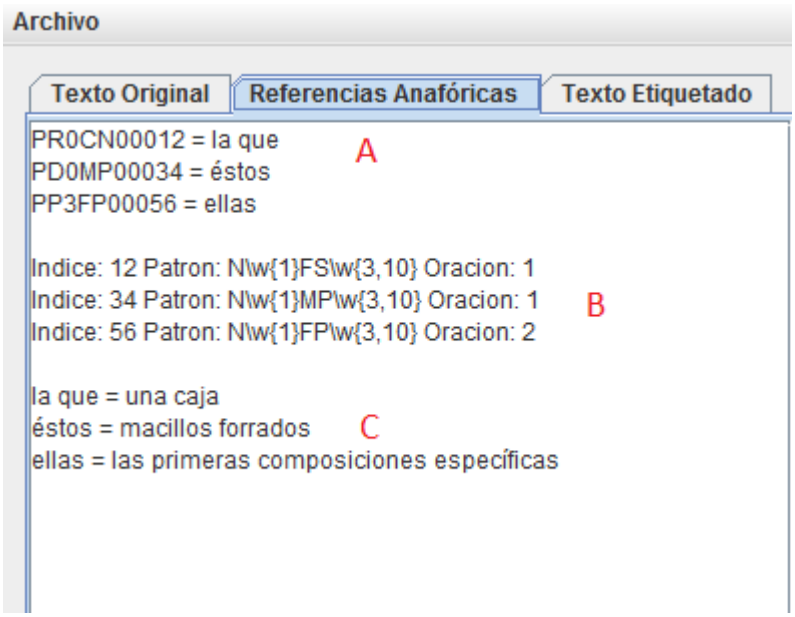
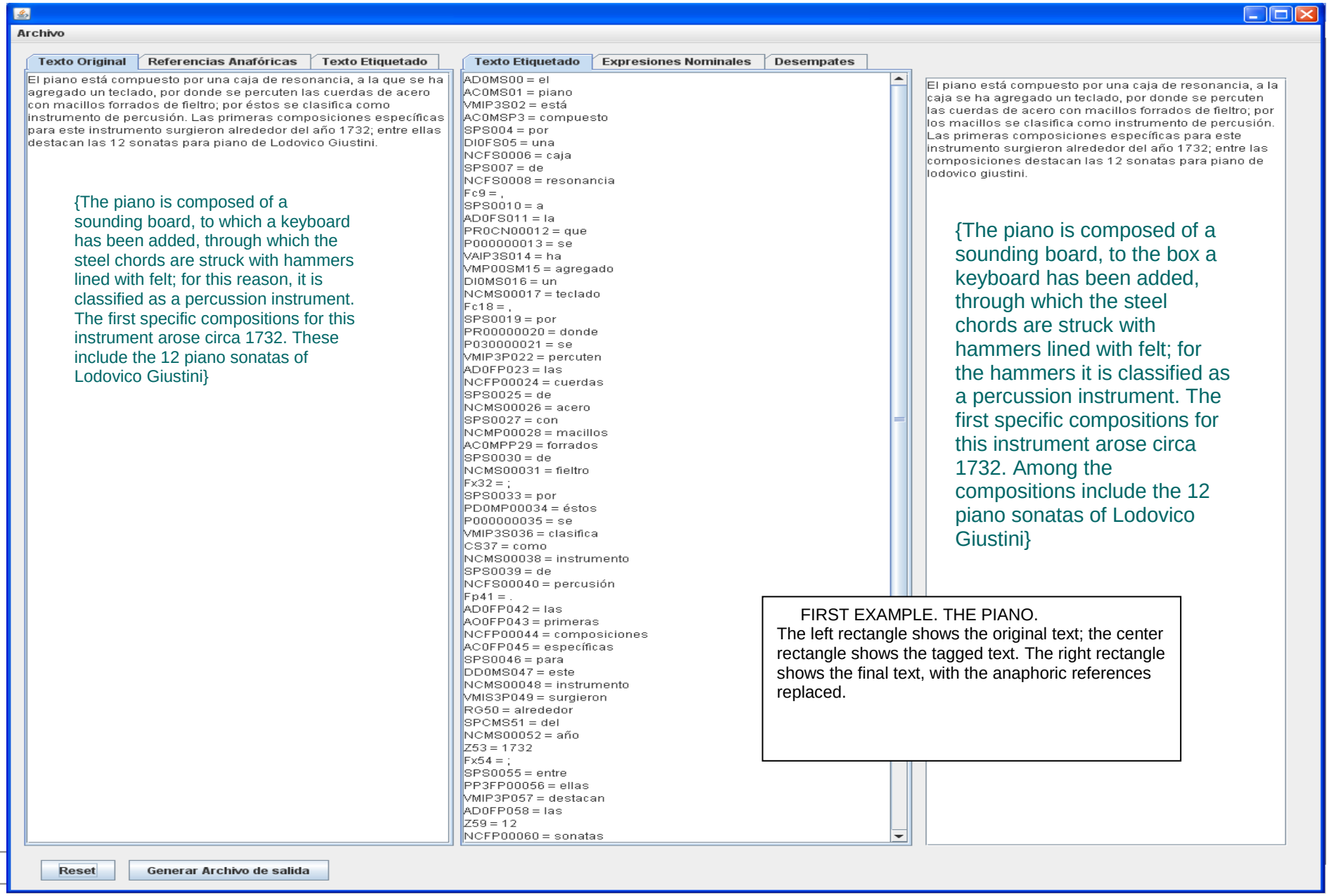


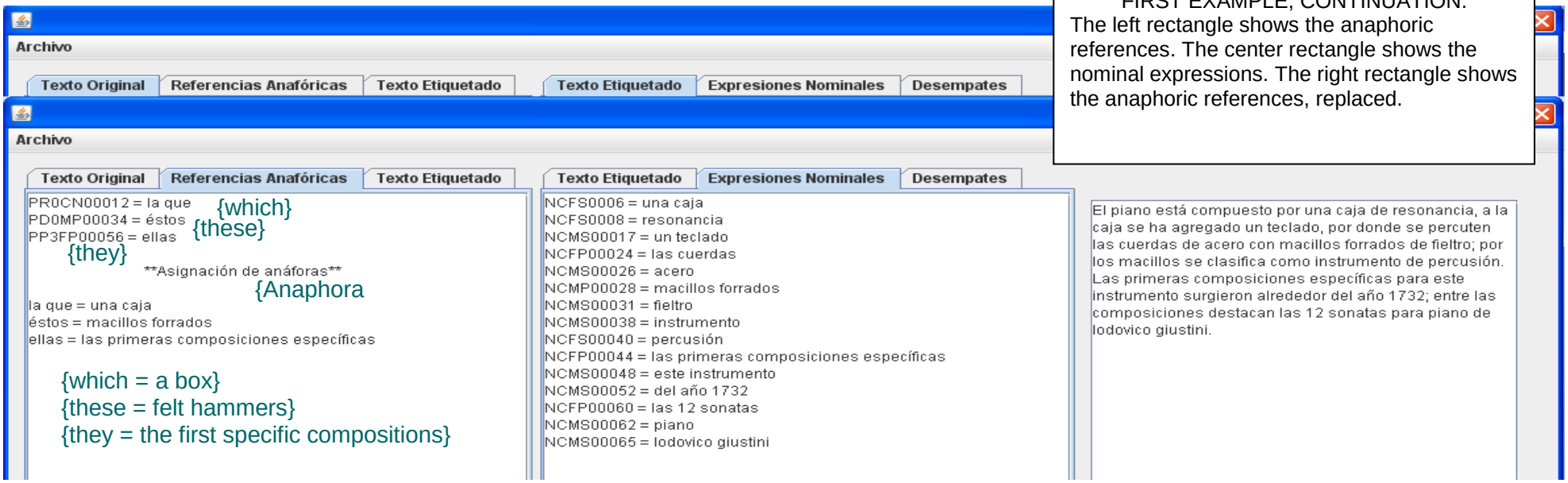
Figure 4 – Tag: Anaphoric references

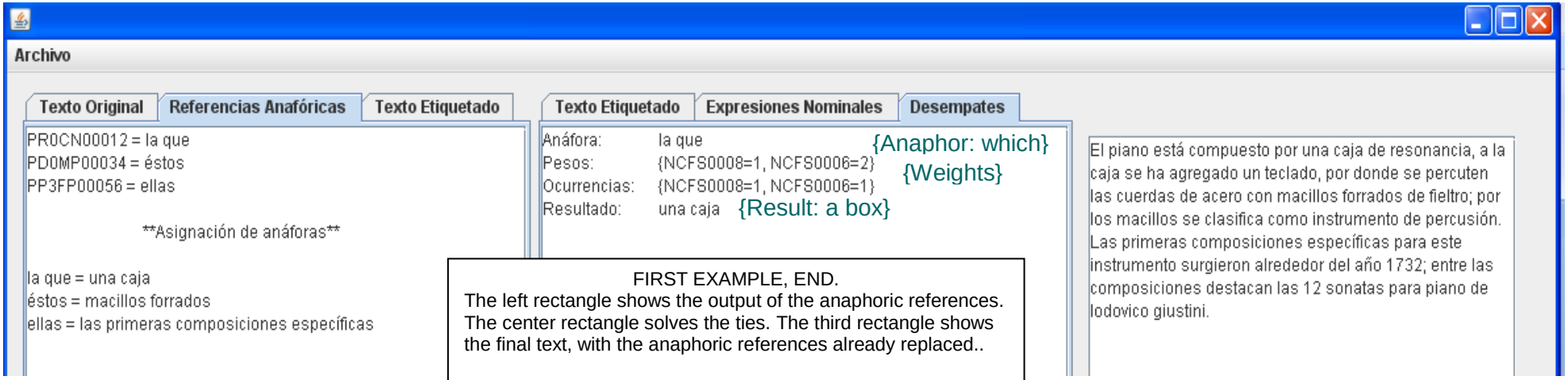
Annex C. Other examples



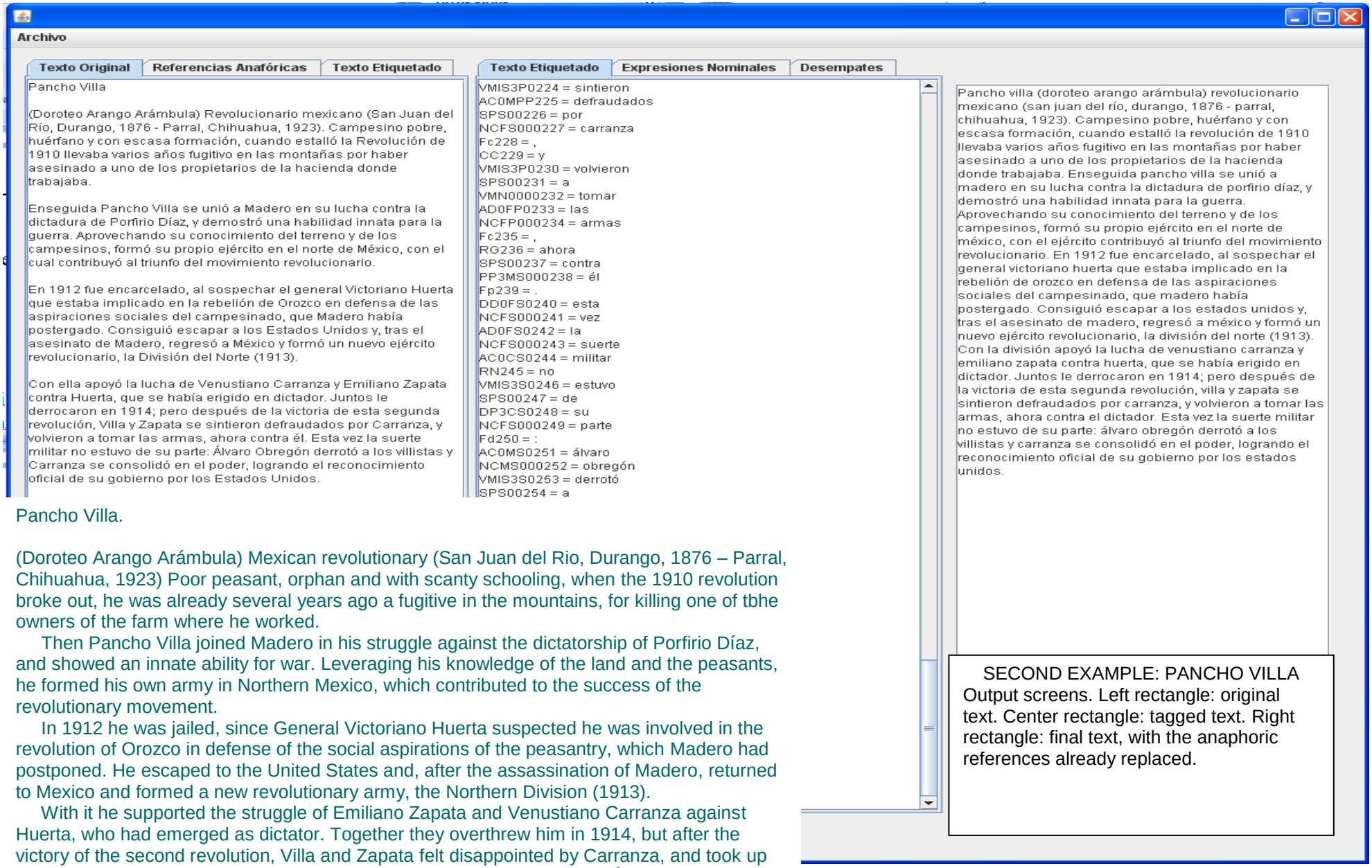
FIRST EXAMPLE. THE PIANO.
The left rectangle shows the original text; the center rectangle shows the tagged text. The right rectangle shows the final text, with the anaphoric references replaced.

FIRST EXAMPLE, CONTINUATION.
The left rectangle shows the anaphoric references. The center rectangle shows the nominal expressions. The right rectangle shows the anaphoric references, replaced.





FIRST EXAMPLE, END.
The left rectangle shows the output of the anaphoric references. The center rectangle solves the ties. The third rectangle shows the final text, with the anaphoric references already replaced..



Pancho Villa.

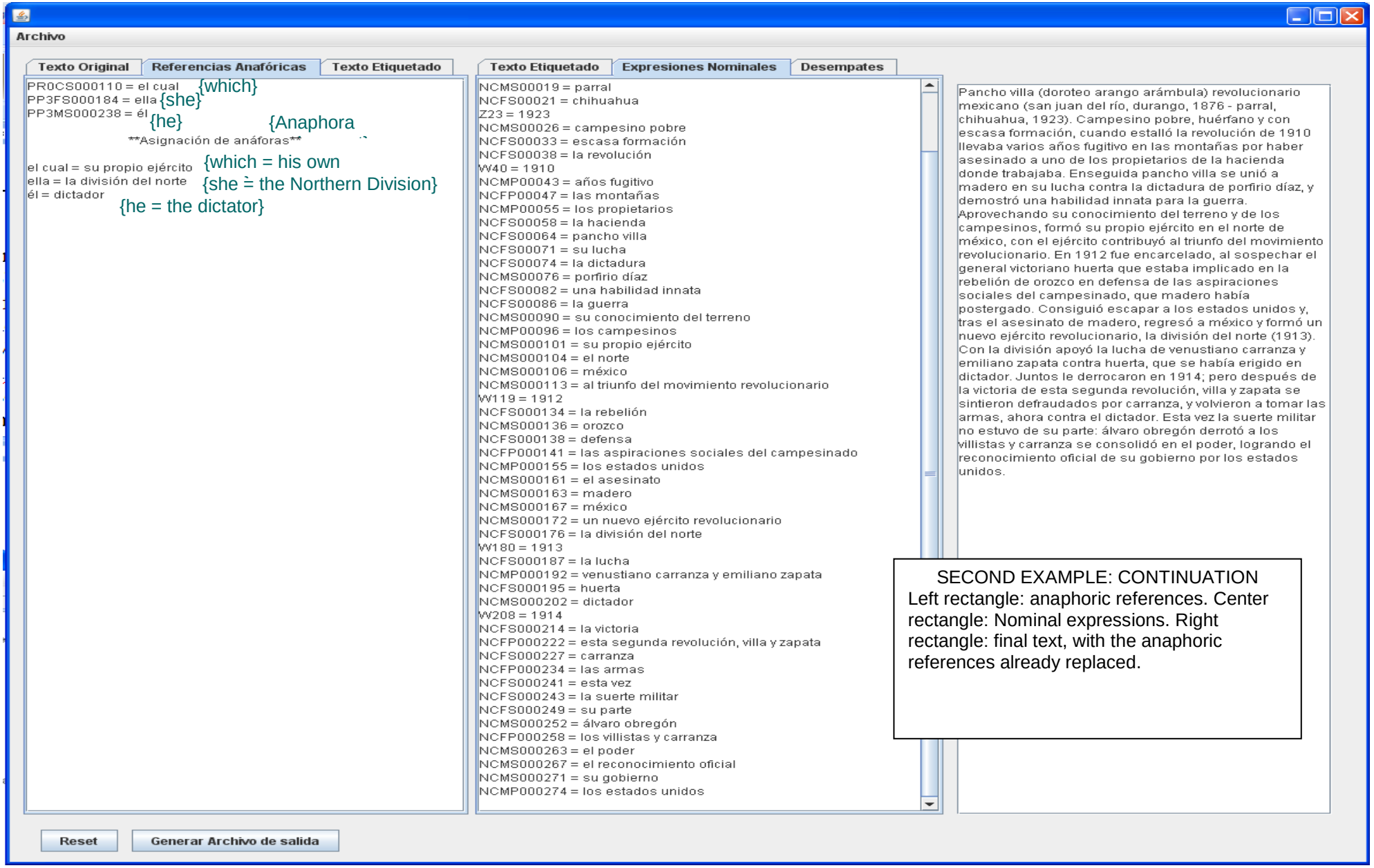
(Doroteo Arango Arámbula) Mexican revolutionary (San Juan del Rio, Durango, 1876 – Parral, Chihuahua, 1923) Poor peasant, orphan and with scanty schooling, when the 1910 revolution broke out, he was already several years ago a fugitive in the mountains, for killing one of tbhe owners of the farm where he worked.

Then Pancho Villa joined Madero in his struggle against the dictatorship of Porfirio Díaz, and showed an innate ability for war. Leveraging his knowledge of the land and the peasants, he formed his own army in Northern Mexico, which contributed to the success of the revolutionary movement.

In 1912 he was jailed, since General Victoriano Huerta suspected he was involved in the revolution of Orozco in defense of the social aspirations of the peasantry, which Madero had postponed. He escaped to the United States and, after the assassination of Madero, returned to Mexico and formed a new revolutionary army, the Northern Division (1913).

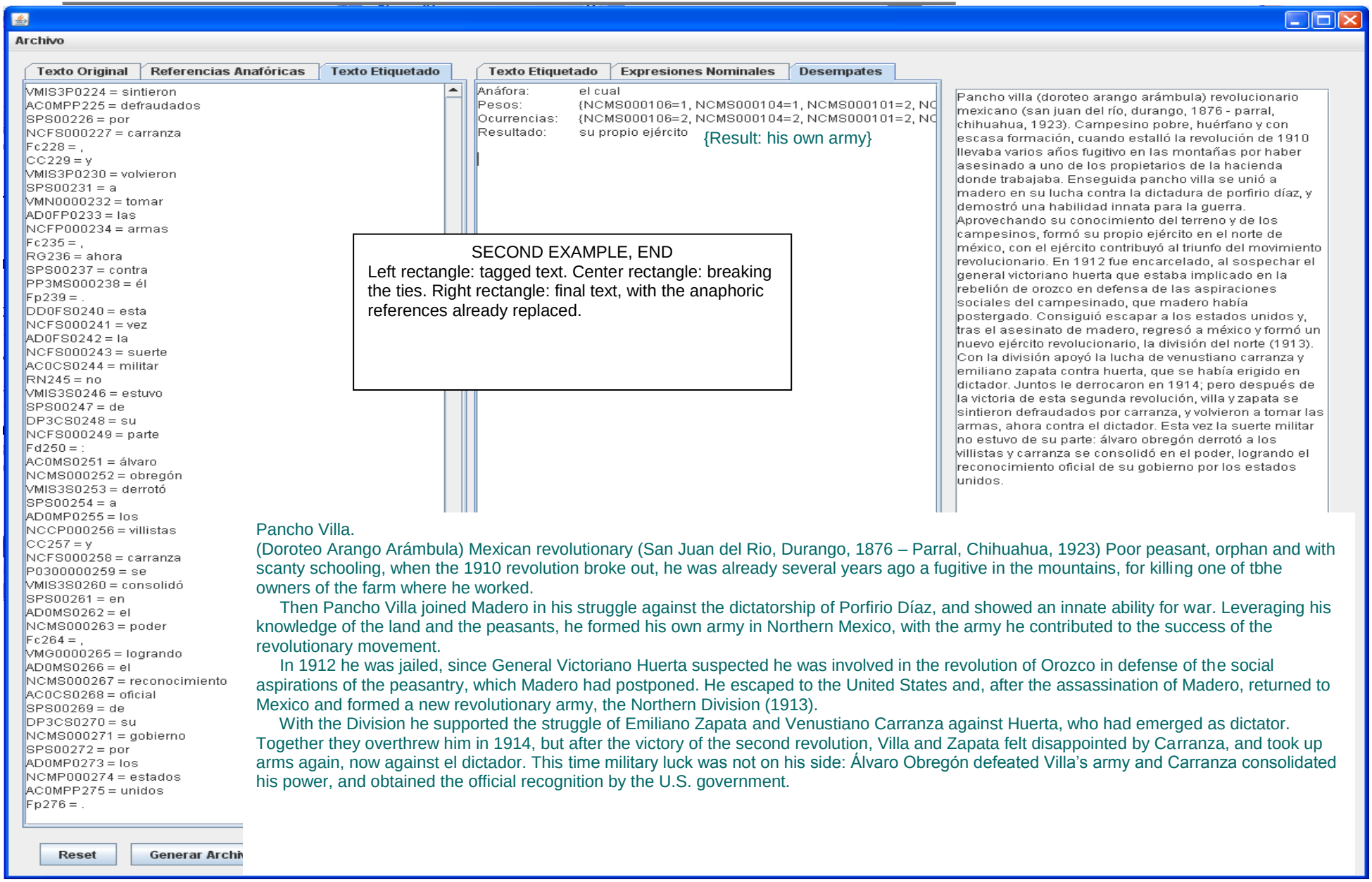
With it he supported the struggle of Emiliano Zapata and Venustiano Carranza against Huerta, who had emerged as dictator. Together they overthrew him in 1914, but after the victory of the second revolution, Villa and Zapata felt disappointed by Carranza, and took up arms again, now against him. This time military luck was not on his side: Álvaro Obregón defeated Villa's army and Carranza consolidated his power, and obtained the official

SECOND EXAMPLE: PANCHO VILLA
Output screens. Left rectangle: original text. Center rectangle: tagged text. Right rectangle: final text, with the anaphoric references already replaced.



SECOND EXAMPLE: CONTINUATION

Left rectangle: anaphoric references. Center rectangle: Nominal expressions. Right rectangle: final text, with the anaphoric references already replaced.



SECOND EXAMPLE, END

Left rectangle: tagged text. Center rectangle: breaking the ties. Right rectangle: final text, with the anaphoric references already replaced.

Pancho Villa.

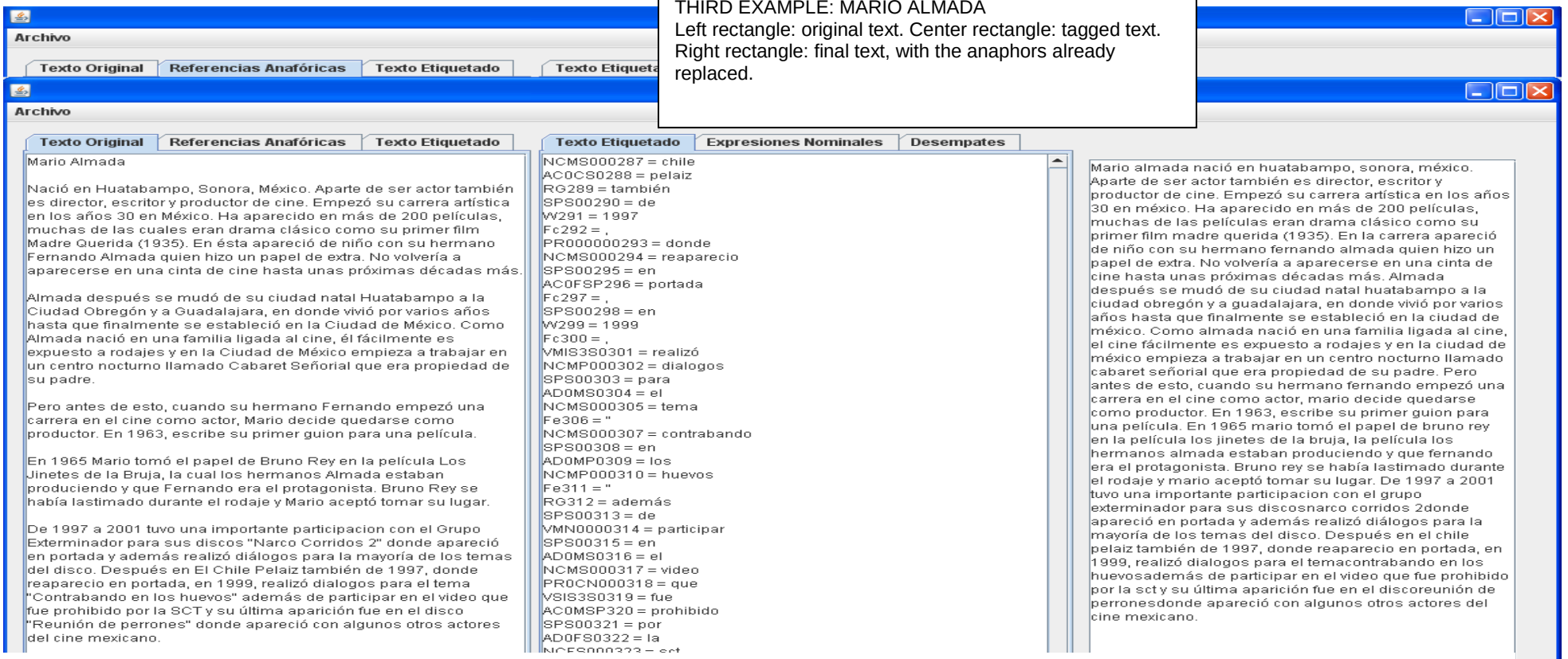
(Doroteo Arango Arámbula) Mexican revolutionary (San Juan del Rio, Durango, 1876 – Parral, Chihuahua, 1923) Poor peasant, orphan and with scanty schooling, when the 1910 revolution broke out, he was already several years ago a fugitive in the mountains, for killing one of tbhe owners of the farm where he worked.

Then Pancho Villa joined Madero in his struggle against the dictatorship of Porfirio Díaz, and showed an innate ability for war. Leveraging his knowledge of the land and the peasants, he formed his own army in Northern Mexico, with the army he contributed to the success of the revolutionary movement.

In 1912 he was jailed, since General Victoriano Huerta suspected he was involved in the revolution of Orozco in defense of the social aspirations of the peasantry, which Madero had postponed. He escaped to the United States and, after the assassination of Madero, returned to Mexico and formed a new revolutionary army, the Northern Division (1913).

With the Division he supported the struggle of Emiliano Zapata and Venustiano Carranza against Huerta, who had emerged as dictator. Together they overthrew him in 1914, but after the victory of the second revolution, Villa and Zapata felt disappointed by Carranza, and took up arms again, now against el dictador. This time military luck was not on his side: Álvaro Obregón defeated Villa's army and Carranza consolidated his power, and obtained the official recognition by the U.S. government.

THIRD EXAMPLE: MARIO ALMADA
Left rectangle: original text. Center rectangle: tagged text. Right rectangle: final text, with the anaphors already replaced.



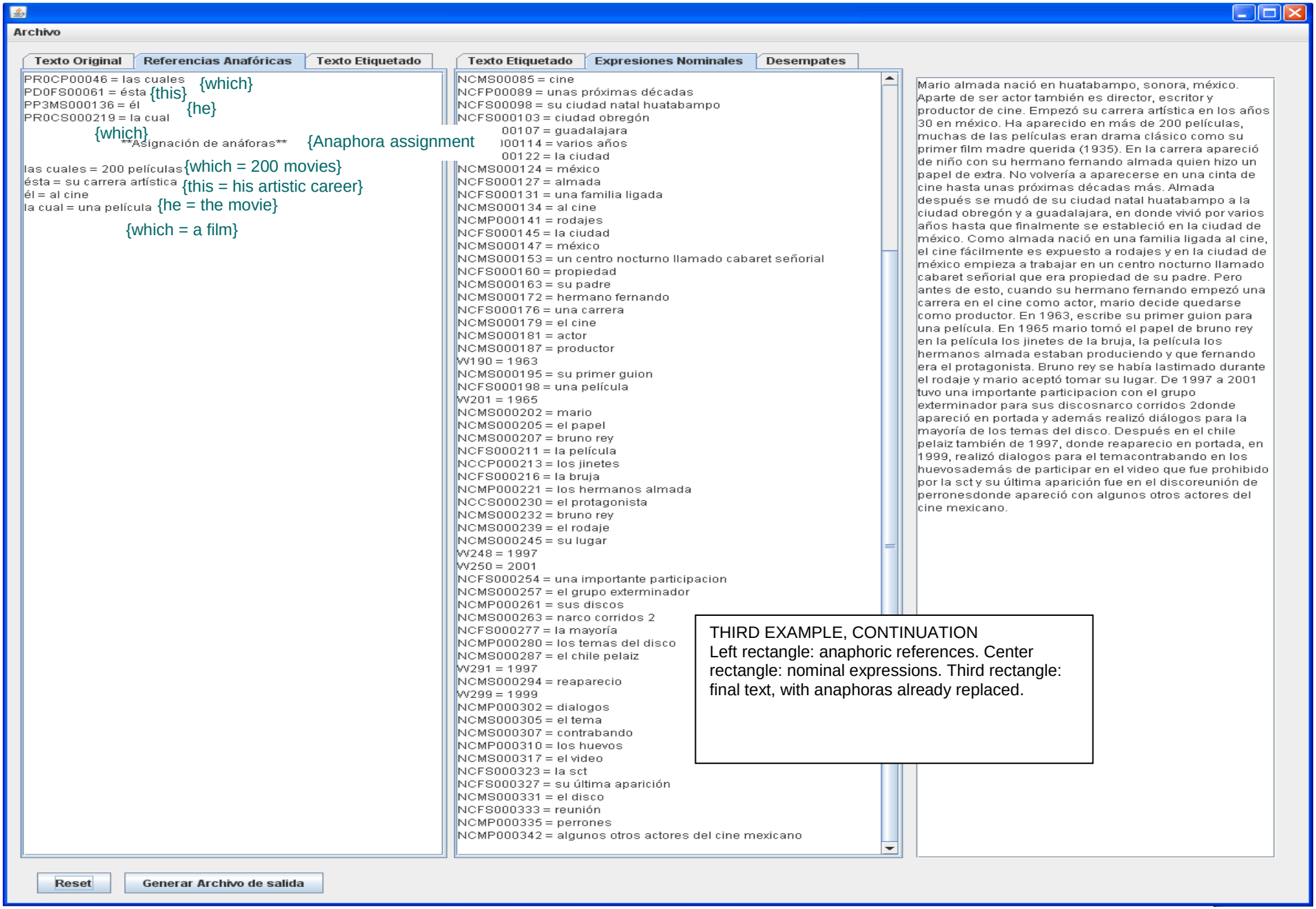
Mario Almada.- Born in Huatabampo, Sonora, Mexico. In addition of being a great actor, he is also director, writer and movie producer. He started his artistic career in the 1930's in Mexico. He has appeared in more than 200 movies, many of which were classic drama, like his first film *Madre Querida* (1935). In this he appeared as a child with his brother Fernando Almada, who performed as an extra. He would not appear again in a movie until several decades later.

Almada moved from his native Huatabampo to Ciudad Obregón and Guadalajara, where he lived for several years until he finally settled in Mexico City. Since Almada was born in a family linked to films, he is easily exposed to filming and in Mexico City he starts working in a night club called *Cabaret Señorial* which belonged to his father.

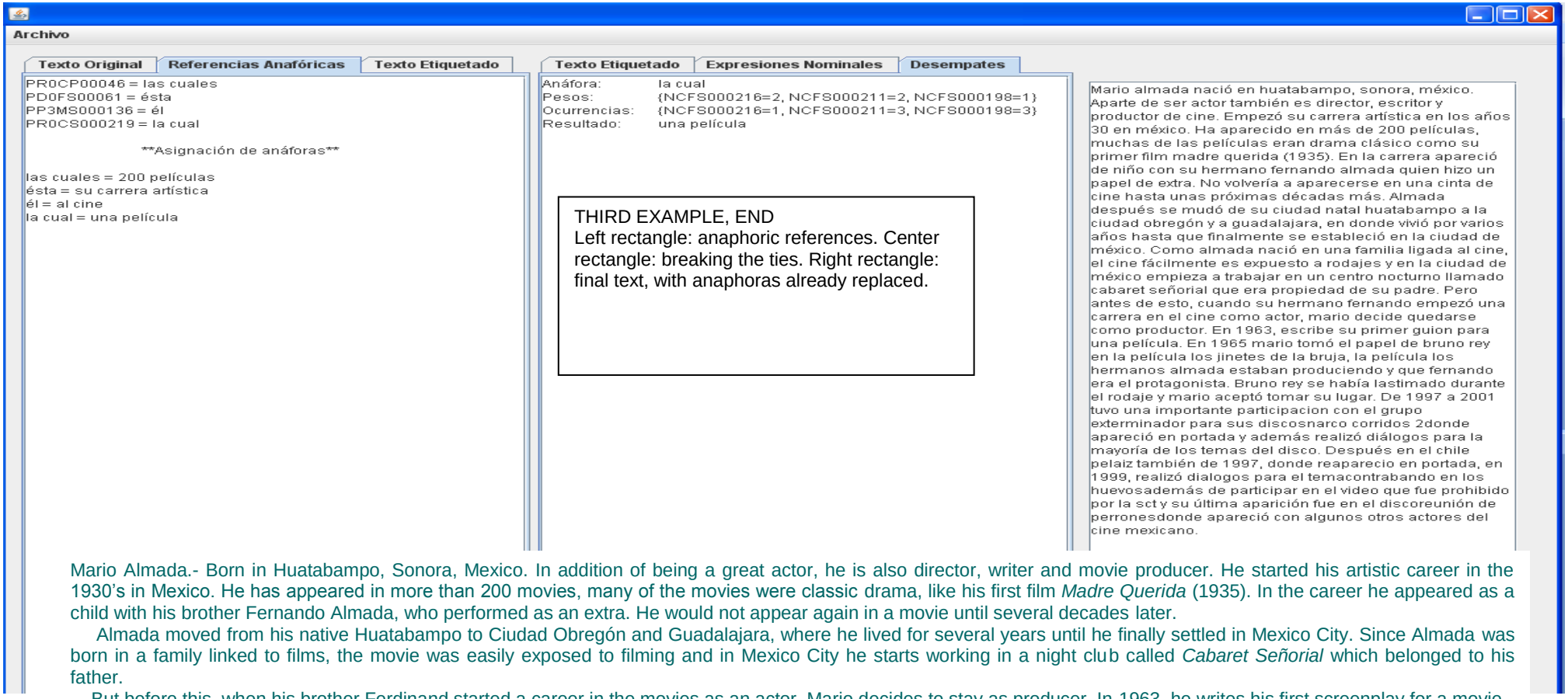
But before this, when his brother Ferdinand started a career in the movies as an actor, Mario decides to stay as producer. In 1963, he writes his first screenplay for a movie.

In 1965 Mario performed the role of Bruno the King in the movie Riders of the Witch, which the Almada brothers were producing, and where Fernando was the first actor. Bruno the King became injured during the filming and Mario accepted to take his place.

From 1997 to 2001 he had an important participation with the group The Exterminator for their records "Narco Corridos 2" where he appeared in the cover, and he also conducted dialogues for most tracks in the album. After that, in *El Chile Peláiz* also from 1997, where he reappeared in the cover, in 1999 he performed dialogues for the theme "Smuggling in the eggs" in addition to taking part in the video that was prohibited by SCT, and his last appearance was in the album "Gathering of dogs" where he appeared with some other Mexican actors.



THIRD EXAMPLE, CONTINUATION
Left rectangle: anaphoric references. Center rectangle: nominal expressions. Third rectangle: final text, with anaphoras already replaced.



THIRD EXAMPLE, END
Left rectangle: anaphoric references. Center rectangle: breaking the ties. Right rectangle: final text, with anaphoras already replaced.

Mario Almada.- Born in Huatabampo, Sonora, Mexico. In addition of being a great actor, he is also director, writer and movie producer. He started his artistic career in the 1930's in Mexico. He has appeared in more than 200 movies, many of the movies were classic drama, like his first film *Madre Querida* (1935). In the career he appeared as a child with his brother Fernando Almada, who performed as an extra. He would not appear again in a movie until several decades later.

Almada moved from his native Huatabampo to Ciudad Obregón and Guadalajara, where he lived for several years until he finally settled in Mexico City. Since Almada was born in a family linked to films, the movie was easily exposed to filming and in Mexico City he starts working in a night club called *Cabaret Señorial* which belonged to his father.

But before this, when his brother Ferdinand started a career in the movies as an actor, Mario decides to stay as producer. In 1963, he writes his first screenplay for a movie.

In 1965 Mario performed the role of Bruno the King in the movie Riders of the Witch, the movie the Almada brothers were producing, and where Fernando was the first actor. Bruno the King became injured during the filming and Mario accepted to take his place.

From 1997 to 2001 he had an important participation with the group The Exterminator for their records "Narco Corridos 2" where he appeared in the cover, and he also conducted dialogues for most tracks in the album. After that, in *El Chile Peláiz* also from 1997, where he reappeared in the cover, in 1999 he performed dialogues for the theme "Smuggling in the eggs" in addition to taking part in the video that was prohibited by SCT, and his last appearance was in the album "Gathering of dogs" where he appeared with some other Mexican actors.

For lack of space the results of the documents "El burro flautista" {The Flutist Donkey} [Online source 2] y "Batalla de Puebla" {Puebla's battle} [Online source 3] are not shown.