The Functionality of a Tool Bar for Postedition in Machine Translation between Languages with Linguistic Interference: the Spanish-Galician Case

Antonio Sayáns Gómez and Elena Villar Conde

Centro Ramón Piñeiro para a Investigación en Humanidades
Estrada Santiago-Noia, km. 3, A Barcia- Santiago de Compostela (Galicia-Spain)
{asayans, evillar}@cirp.es

Abstract

The Department of Linguistics of the *Centro Ramón Piñeiro para a Investigación en Humanidades (C.R.P.I.H.)*, headed by Professor Guillermo Rojo, has developed *Es-Ga*, a machine translation system based on the Metal system which at the present time translates from Spanish into Galician in .rtf, .txt and .html formats. It also contains a number of programmes whose function is to deformat documents that are then translated and, once this process has finished, to reconstruct their original format. The system has a tool bar with linguistic information designed for MS-WORD, the functionality and functioning of which has proven unquestionable as an aid to the posteditor in a context of linguistic interference between two intercomprehensible languages.

1 Introduction

Machine translation between two genetically close languages has certain characteristics that make it distinct from machine translation between less related languages. This is the case of translation between Galician and Spanish, two Romanic languages spoken in Galicia (a region in North West Spain). Their lexical and morphosyntactic similarities potentiate certain linguistic phenomena, also influenced by a situation of diglossia that has marked a linguistic context characterised by the greater prestige of Spanish with respect to Galician.

All this, together with a high degree of intercomprehensibility[1], has been a breeding ground for the linguistic interferences that we will delve into later.

Thus, in elaborating the translation system *Es-Ga*[2], the syntactic and morphological structure similarity between Spanish and Galician has been taken profit of. However, the profit is limited, since the system programmer has the difficult task of establishing the contextual background for the Spanish structures. In many cases, this difficulty is determined by an ability to tackle the relevant linguistic structures for the transference from one language to the other.

Bearing all this in mind, the convenience was considered of elaborating a device that should draw attention upon the linguistic structures at issue. Such a device would contribute to the structures receiving special attention by the posteditor. Let us take into account the fact that she herself may have internalised many of those structures and may hence, after a superficial review, accept an incorrect translation.

In this way the idea arose of a tool bar[3] designed for MS-WORD fulfilling the focusing function just alluded to, offering the posteditor specific linguistic information.

In this work we intend to show its usefulness as an aid to the posteditor in a context of linguistic interference between two intercomprehensible languages.

2 Linguistic Interferences in Machine Translation

[1] Intercomprehensibility leads to the assimilation in Galician of many interferences from Spanish, the Galician speakers not being aware of usage of elements from another language.

[2] A transference system was selected on the grounds that this would produce the best results since the languages involved are genetically related and their syntactic and grammatical structures are similar.

[3] Devised with the collaboration of the firm Sail Labs, already disappeared.

Already in 1953 Uriel Weinrich stated in *Languages in contact* that "the term interference implies the rearrangement of patterns that result from the introduction of foreign elements into the more highly structured domains of language, such as the bulk of the phonemic system, a large part of the morphology and syntax, and some areas of the vocabulary (kinship, color, weather, etc.)"[4]. These interferences may be bidirectional, but, in a linguistic context where a prestigious language (Spanish in this case) coexists with a language that is not so highly valued (Galician in this case), the interferences are socially determined. Thus, assimilation of morphosyntactic and lexical structures in the "low-category" language is seen as positive, and, conversely, interferences in the "high-category" language from the "low-category" language will result in a certain loss of prestige. This explains why Galician speakers continuously assimilate lexical and grammatical elements from Spanish.

Linguistic interferences may be divided into two classes: lexical interferences and morphosyntactic interferences.

a) Lexical interferences: These are most easily recognisable. They affect the lexicon, and we may thus find pure Spanishisms (*silla* instead of *cadeira* 'chair'), ultracorrections (*brilar* instead of *brillar* 'shine'), adaptations of Spanishims (*basureiro* instead of *vertedoiro do lixo* 'rubbish dump') and semantic calques. This class of interferences affects in a special way certain semantic fields such as family relations (*abuelo* instead of *avó* 'grandfather'), colours (*amarillo* instead of *amarelo* 'yellow'), religion (*Dios* instead of *Deus* 'God')... But, in a language like Galician, currently undergoing a normalisation process, technological vocabulary is the type of vocabulary in which this class of interferences is most difficult to recognise. This is due to the fact that a large proportion of this vocabulary was introduced in Galician from other languages through Spanish and hence the Galician version either is not settled or even is completely unknown to the posteditor (*cf. llanta* instead of *lamia* 'rim' –Cars–).

b) Morphosyntactic interferences: Galician speakers assimilate this class of interferences in a higher degree as these come to fit naturally into their grammar. They occur at a deeper level than lexical interferences. Properly morphological are calques in gender (feminine *a leite* instead of masculine *o leite* 'milk' –with the feminine determiner, as, in Galician, uncountable nouns typically take a determiner–) or in the formation of plural (singular *lapis* 'pencil' > plural *lápices* instead of *lapis* 'pencils'), derivation processes (*pobreciño* instead of *pobriño* 'poor' with a diminutive denoting affection), or word formation (*aterrizar* instead of *aterrar* 'land' –verb–). Properly syntactic are incorrectly used prepositions (*atender ós feridos* –'see' + contraction of preposition and determiner + 'injured'– instead of *atender os feridos* –'see' + determiner + 'injured'– 'see the injured'), incorrect placing of pronouns (*me dixo* instead of *díxome* 'he/she told me'), use of compound tenses –non existent in normative Galician– (*había feito* por *fixera* 'he/she had done'), reflexive verbs –non existent in normative Galician– (*sentouse no chan* instead of *sentou no chan* 'he/she sat on the floor')...

Interferences, besides the structural similarity between the two languages at issue, constitute a handicap in the task of designing grammars in the machine translation system. Furthermore, this structural similarity also influenced our translation work because of the difficulties to precisely codify, in the translation system, relevant clues to identify contextual aspects so as to prevent the convergence of some divergent structures in source and target language.

On the one hand, syntactic and morphological structural similarity may be said to constitute an advantage to the linguist in the task of reordering

[4] Weinreich, U. (ed.) (1968). *Languages in contact*, The Hague/Paris: Mouton, p.1.

the generation grammar[5] of the target language, and to the posteditor as she will find the structures familiar, with grammatical devices which are very close in the two languages. But, on the other hand, that same similarity turns into a disadvantage when, changing from one language to the other, morphosyntactic structures vary, and, at the same time, there are difficulties in contextualisation, this latter being essential in providing the system with specific and precise clues that make possible the production of an appropriate translation. It corresponds to the posteditor the task of making a decision wherever the system has not been able, by itself, to find the correct option. But the posteditor's assimilation of interferences might prevent the appropriate completion of this task. A tool is then needed that informs her of those cases where the system cannot guarantee the appropriate translation due to a lack of context or of vocabulary not found in its lexicons[6].

It is on the basis of this need that the *Es-Ga* tool bar has been devised.

3 The *Es-Ga* Tool Bar

In principle, the *Es-Ga* tool bar was devised to draw attention upon the lack of lexicon in the dictionaries and to disambiguate certain expressions by providing a range of options from which the posteditor should ultimately decide.

But, needless to say, this tool bar may be very helpful in the treatment of linguistic interferences from Spanish to Galician. In this way, much importance has been attributed to the fact that the posteditor is working with closely related languages in which she is highly competent; and to the possibility that –due to the translated text's intelligibity- the posteditor does not notice these interference errors, which she may well have internalised, and she overlooks them in a superficial review.

Hence the functionality of this tool bar, which, predicting the possibility of error, present the posteditor with a range of options.

The *Es-Ga* tool bar works with .txt and .rtf formats, which means it is incorporated exclusively in MS-WORD. This responds to the actual user's needs, since translation is most common in those formats. The bar consists of a group of eight macros executable through their respective commands which are represented visually in the bar buttons.

3.1 Tagging

When translating, the system incorporates a tagging to the document. The tags may be decodified in postedition with the tool bar. They affect, on the one hand, words for which the system does not find equivalence in the different dictionaries and, on the other hand, structures presenting some linguistic problem that makes their translation difficult. Once the text has been taken from the system's outbox and the document opened with MS-WORD, one of de macros will make it possible to process the tags[7], incorporating the relevant comment for each of the tags and applying an ascendant numbering.

3.2 Use of the *Es-Ga* Tool Bar

In what follows we are going to explain how the *Es-Ga* tool bar works. The posteditor, once the bar has been activated, has the following buttons at her disposal:

- *Depura_Documento* ('Refines_Document')
- *Procesa_Marcas* ('Processes_Marks')
- *Recupera_Orixinal* ('Recovers_Original')
- *Posible_CD* ('Possible_Direct Object')
- *Tempo_Verbal* ('Verb_Tense')
- *Elimina_Comentarios_Ling* ('Eliminates_Linguistic_Com ments')

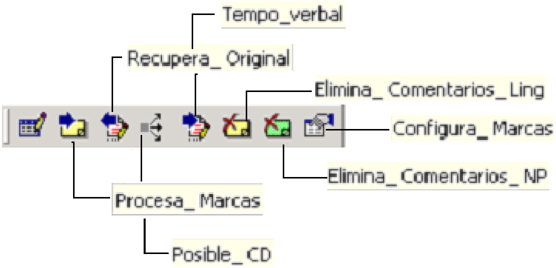
[5] *Es-Ga* contains three grammars: a grammar for the syntactic analysis of Spanish, provided by the, already disappeared, Catalan firm Incyta; a grammar for the transference from Spanish to Galician, turning the linguistic structures of the source language into linguistic structures of the target language; and a grammar for the generation of Galician, dealing with the final result of the Galician text.

[6] The system has three lexicons: monolingual Spanish, devised by the *CRPIH* and Incyta (49,000 entries) and monolingual Galician (47,500 entries) contain syntactic, morphological and semantic information for each entry; bilingual Spanish-Galician (47,200 entries) provides word-to-word correspondences as well as scrutinising possible contextual aspects of the Spanish entry in order to achieve the best semantic match.

[7] The processing of the tags consists of eliminating the mark visually and highlighting the text affected by the tag.

- *Elimina_Comentarios_NP* ('Eliminates_Proper Noun_Comments')

Configura_Marcas ('Configures_Marks')



Acording to their functionality, these buttons may be classified in:

a) Text processing: *Depura_Documento*, *Procesa_Marcas*, *Recupera_Orixinal*, *Elimina_Comentarios_Ling*, *Configura_Marcas* and *Elimina_Comentarios_NP*.

b) Linguistic information: *Posible_CD* and *Tempo_Verbal*.

A step-by-step explanation follows in order to illustrate how the bar functions when applied to a translation.

The text opened and refined, the marks must be configured. Pressing the appropriate button, a dialogue window is opened in which the originaly activated marks appear. These are the following:

- *GA-falta no léxico galego* ('GA-absent from the Galician lexicon'): with this option, all the lexicon the system has not incorporated in the monolingual Galician dictionary will appear highlighted in brown.

E.g.: in a Spanish text with the toponym *Cisjordania* ('The West Bank'), this toponym would come out from the translation machine tagged as follows: <GA>*Cisjordania*</GA>; after the processing of the marks, Galician *Cisxordania* will appear, highlighted in brown, which indicates that the toponym is present in the bilingual dictionary but absent from the monolingual Galician dictionary.

- *TR-falta no léxico de transferencia* ('TR-absent from the transference lexicon'): with this option all the lexicon that the system has not incorporated in the bilingual dictionary will appear highlighted in olive green.

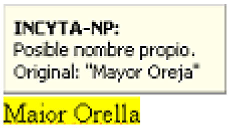
E.g..: *declaradamente* ('openly') would come out tagged as <TR>*declaradamente*</TR> and processed by the bar as *declaradamente*.

- *ES-falta no léxico castelán* ('ES-absent from the Spanish lexicon'): with this option all the lexicon that the system has not incorporated in the monolingual Spanish dictionary will appear highlighted in blue.

E.g.: the foreign *yihad* ('yihad') absent from the Spanish lexicon would be tagged by the system as <ES>*yihad*</ES> and the posteditor would be able to see it highlighted in blue after the marks have been processed: *yihad*.

- *NP-posible nome propio* ('NP-possible proper noun'): due to the possibility of an incorrect translation of proper nouns, a method was sought that should signal words with semantic meaning that might be functioning as proper nouns when beginning with capital letters.

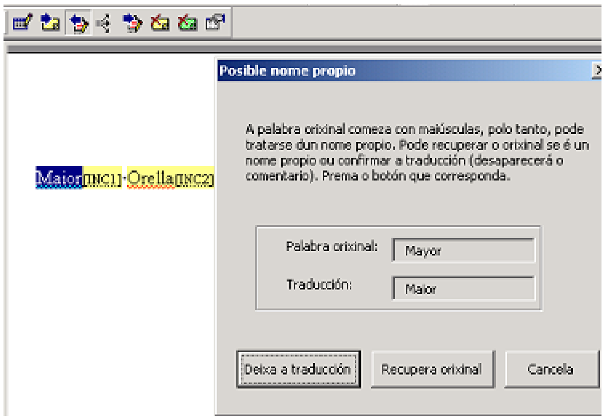
E.g.: the surnames of a well-known Spanish politician, *Mayor Oreja*, are potentially translatable as *Maior Orella* ('Bigger Ear'), the system warns us that it could be a proper noun by means of the tagging <NP>*Maior*</NP><SR>*Mayor*</SR> <NP>*Orella*</NP> <SR>*Oreja*</SR> and, after this tag is configured, highlights the expression in a yellow square, and provides a comment:



INCYTA-NP:
Posible nombre propio.
Original: "Mayor Oreja"

Maior Orella

The button *Recupera_Orixinal* ('Recovers_Original') offers the possibility to leave a word untranslated.



Maior[INC1] Orella[INC2]

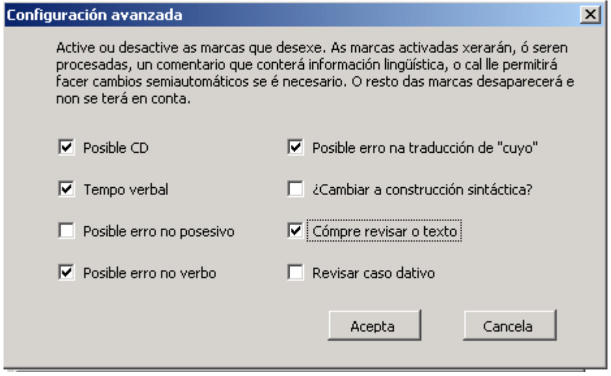
Posible nome propio

A palabra orixinal comeza con maiúsculas, polo tanto, pode tratarse dun nome propio. Pode recuperar o orixinal se é un nome propio ou confirmar a traducción (desaparecerá o comentario). Prema o botón que corresponda.

Palabra orixinal: Mayor

Traducción: Maior

The posteditor has the possibility of configuring other marks with linguistic information concerning the morphosyntactic structure. For this, she has to press the button *Máis...* ('More...') in the window *Configuración de marcas de traducción* ('Configuration of translation marks**'**) and thus access a new window where she can select among eight new marks. These new marks are the following: *Posible CD* ('Possible Direct Object'), *Tempo verbal* ('Verb tense'), *Posible erro no posesivo* ('Possible error in the possessive adjective'), *Posible erro no verbo* ('Possible error in the verb'), *Posible erro na tradución de "cuyo"* ('Possible error in the translation of "cuyo"-"whose"'), *¿Cambiar construcción sintáctica?* ('Change sintactic construction?'), *Cómpre revisar o texto* ('Text review needed') and *Revisar o dativo* ('Review the dative case').



Configuración avanzada

Active ou desactive as marcas que desexe. As marcas activadas xerarán, ó seren procesadas, un comentario que conterá información lingüística, o cal lle permitirá facer cambios semiautomáticos se é necesario. O resto das marcas desaparecerá e non se terá en conta.

☑ Posible CD

☑ Tempo verbal

□ Posible erro no posesivo

☑ Posible erro no verbo

☑ Posible erro na tradución de "cuyo"

□ ¿Cambiar a construcción sintáctica?

☑ Cómpre revisar o texto

□ Revisar caso dativo

Some of these marks are particularly useful due to a higher occurrence frequency of the possible error to which they refer.

Now we are going to analyse their usefulness, providing examples:

- *Posible CD* ('Possible Direct Object'): this signals structures translated as Indirect Objects that could in reality be Direct Objects; such a confusion stemming from a widespread grammatical phenomenon in Spanish known as *leísmo*.

This phenomenon consists of the use of *le/les* (respectively third person singular and plural pronoun), traditionally fulfilling the function of Indirect Object, as Direct Object; *le* is thus substituted for *lo/la* (respectively third person singular masculine and feminine pronoun), and *les* for *los/las* (respectively third person plural masculine and feminine pronoun).

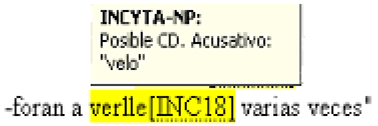
The *R.A.E.* (*Real Academia Española de la Lengua*, '*Royal Academy of the Spanish Language*') accepts *leísmo* only for singular Direct Objects referring to a male person (*lo* > *le*), as in "*le vi en el teatro*" (< "*lo vi en el teatro*") 'I saw him in the theatre', with third person singular pronoun *le*, traditionally used as Indirect Object, functioning as Direct Object, substituting for *lo*, traditionally used as Direct Object.

The phenomenon is not present in the Galician language, and so a problem arises with its translation into this latter language. The basis of the problem may be said to be the double function of the Spanish third person singular pronoun *le*, as Indirect Object (the tradition) and as Direct Object (the modern phenomenon, option accepted by the *R.A.E.* when referring to a male person); or even the double functionality of *le* with other types of reference and of *les*. Thus, the translation of Spanish *le* as Galician third person singular pronoun *lle* for Indirect Object is correct in Galician *sorrinlle* (< Spanish "*le sonreí*") 'I smiled to him'; but the translation of Spanish *le* as Galician *lle* for Indirect Object is incorrect in *e.g.* Galician "**enganeille*" (< Spanish "*le engañé*"), literally *'I betrayed to him' instead of Galician correct "*enganeino*" 'I betrayed him', with the Galician third person singular pronoun -*no* for Direct Object. Had the traditional Spanish version "*lo engañé*" been used, the translation into Galician could have been easily produced, with Spanish *lo* to which Direct Object traditionally corresponds spontaneously translated as Galician -*no* to which Direct Object exclusively corresponds[8].

The corresponding error in translation into Galician is partly due to the impossibility of getting the system to recognise, with a complete effectiveness, when Spanish *le/les* is or is not functioning as a Direct Object.

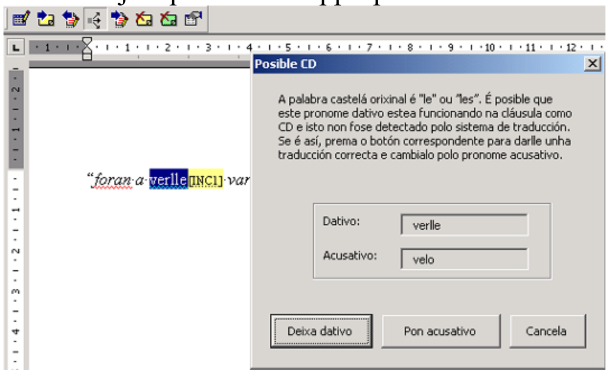
The system does not only produce the tag alluding to a possible error in the pronoun but also incorporates a comment tag to the document: "*foran a* <LL>*verlle*</LL><DM>*velo*</DM> *varias veces*"('They had gone to see him several times').

[8] In fact, Galician has *(-)o/-lo/-no* as allomorphs of the same pronoun, but that is irrelevant for the present purposes.


INCYTA-NP:
Posible CD. Acusativo:
"velo"

-foran a verlle[INC18] varias veces"

When activating the mark *Posible CD* in the configuration button, a button in the tool bar is automatically activated (*Posible_CD*) that offers the posteditor the possibility to choose between the Galician Indirect Object pronoun or the Galician Direct Object pronoun as appropriate.

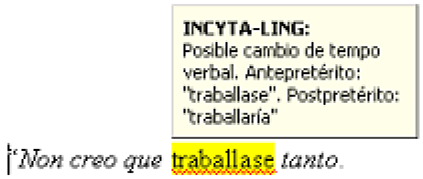


A palabra castelá orixinal é "le" ou "les". É posible que este pronome dativo estea funcionando na cláusula como CD e isto non fose detectado polo sistema de traducción. Se é así, prema o botón correspondente para darlle unha traducción correcta e cambialo polo pronome acusativo.

Dativo: verlle

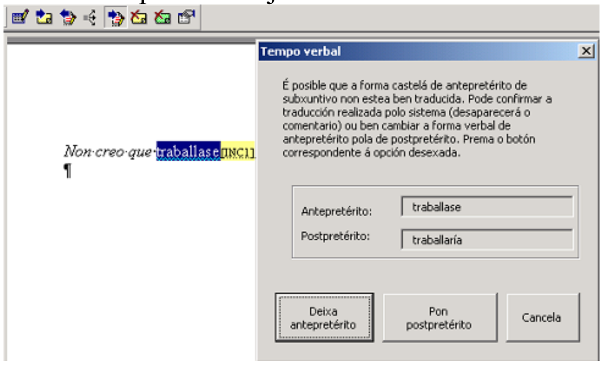
Acusativo: velo

- *Tempo verbal* (‘Verb tense’): the possibility is offered to mark structures translated as imperfect subjunctive that could in reality be conditional, since Spanish imperfect subjunctive, though usually corresponding to Galician imperfect subjunctive, may correspond to Galician conditional. *E.g.*: “*non creo que* <AP>*traballase*</AP><PP>*traballaría*</PP> *tanto* (‘I do not think that he worked so much’).


INCYTA-LING:
Posible cambio de tempo verbal. Antepretérito: "traballase". Postpretérito: "traballaría"

“Non creo que traballase tanto.

Pressing this mark, the button Tempo_Verbal is activated, presenting a dialogue window that offers the posteditor the possibility to chose between imperfect subjunctive and conditional:

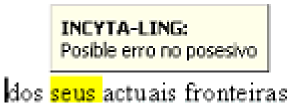


É posible que a forma castelá de antepretérito de subxuntivo non estea ben traducida. Pode confirmar a traducción realizada polo sistema (desaparecerá o comentario) ou ben cambiar a forma verbal de antepretérito pola de postpretérito. Prema o botón correspondente á opción desexada.

Antepretérito: traballase

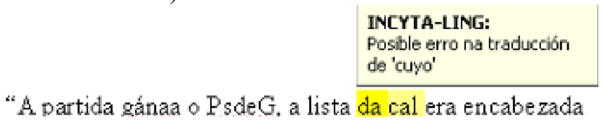
Postpretérito: traballaría

- *Posible erro no posesivo* (‘Possible error in the possessive adjective’): the Spanish possessive adjective referring to a single object possessed is not marked for gender, and ambiguity may exist as to the required gender in Galician. A tag is needed to mark these ambiguous structures. *E.g.*: “*dos* <PO>*seus*</PO> *actuais fronteiras*” (‘of their [masculine] current [feminine] frontiers [feminine]’).


INCYTA-LING:
Posible erro no posesivo

dos seus actuais fronteiras

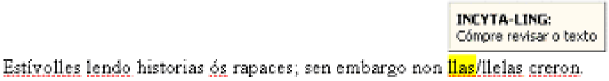
- *Posible erro na traducción de "cuyo"* (‘Possible error in the translation of ‘cuyo’-“whose”’): Spanish relative clauses introduced by the possessive relative adjective *cuyo/cuya/cuyos/cuyas* are highly difficult to translate into Galician. Lacking an exact equivalent for that adjective, Galician generally requires a syntactic restructuring through a Noun Phrase modified by a Prepositional Phrase (e.g.: *el autor cuyo libro* > Galician *o autor o libro do cal*, ‘the author whose book’ –literally ‘the author the book of which’–). The option Noun Phrase + Prepositional Phrase has been chosen to be generated by the system, through the posteditor will have to decide about this and other options (relative clause, Prepositional Phrase…). Furthermore, with Noun Phrase + Prepositional Phrase, the antecedent of the Prepositional Phrase make not be recognised by the system if it is very distant from the Prepositional Phrase or if it is part of a superior structure, and hence the system may produce errors in gender concordance. *E.g.*: “*a partida gánaa o PsdeG, a lista* <CY>*da*</CY> <CY>*cal*</CY> *era encabezada...*” (‘the [feminine] match [feminine] is won by the [masculine] PsdeG [masculine], the list of which was headed…’).


INCYTA-LING:
Posible erro na traducción de 'cuyo'

“A partida gánaa o PsdeG, a lista da cal era encabezada

The antecedent of the possessive relative adjective being masculine *o PsdeG*, the system considers that the antecedent is the subject of the main clause, *i.e.* feminine *a partida*. Even though an appropriate translation comes out in similar examples, the decision corresponds to the posteditor to select among different possibilities avaliable in the Galician language (*e.g.*: *a partida* *gánaa o PsdeG, e a súa lista era encabezada...* 'the match is won by the PsdeG, and its list was headed…'; *a partida gánaa o PsdeG, cunha lista que era encabezada...*, 'the match is won by the PsdeG, with a list that was headed', etc...

- *Cómpre revisar o texto*: when the system has problems to determine the gender or number of an element that has a very distant antecedent, both possibilities (masculine and feminine for gender, singular and plural for number) are offered for the posteditor to decide. *E.g.* "*estívolles lendo historias ós rapaces; sen embargo non* <REV>*llas/llelas*</REV> *creron* ('he/she was reading aloud some stories to the children; however, they did not believe them of him/them')



INCYTA-LING:
Cómpre revisar o texto

Estívolles lendo historias ós rapaces; sen embargo non llas/llelas creron.

After the configuration of the marks that the posteditor has considered to be relevant, these have to be processed. To do that, it is necessary to press the bar button (*Procesa_Marca*), which eliminates the tags visually, with the result that the text is marked in two different ways. In order to mark a lack in the lexicon, a colour gradation is used. In order to mark difficulties for analysis, the text is shaded in yellow. For these latter marks, as the cursor is situated on the marked word(s), a comment appears referring to the type of difficulty involved.

Besides the possibility of recovering information from the original text, the posteditor has two buttons at her disposal for eliminating two classes of tags, if not interested in seeing them. These are the tag referring to a possible proper noun (*NP*) and all the tags related to linguistic issues except for those referring to a lack of lexicon in the different dictionaries.

4 Conclusion

With this work, we hope to have shown that machine translation between genetically related and close languages is not so simple as it might seem at first sight, presenting problems specific to it which are materialised in linguistic interferences and ambiguities. Our suggestion for tackling these disadvantages is the *Es-Ga* tool bar. The textual marks that it offers make it possible to clearly detect which the properly linguistic difficulties in the text are, which difficulties are due to a lack of lexicon in the different dictionaries, and, finally, which translated words may correspond to proper nouns and should, if they do, be kept in their original form. All these options offered by the *Es-Ga* tool bar allow a greater dynamism in work and increase the guarantees of a correct result as they are devised to target the real problems of translation from Spanish into Galician. Needless to say, the posteditor has always the possibility of doing without the tool if she considers it appropriate.

5 Bibliographical References

DIZ, I. (2001). "The importance of MT for the survival of minority languages: Spanish-Galician MT system". In *Proceedings of the MT SUMMIT VIII, Machine Translation in the Information Age*, Santiago de Compostela, Spain.

DIZ, I., and L. MARTÍNEZ (2000). "The Spanish-Galician and Galician-Spanish MT system: How to re-use the existing Galician resources to develop a robust MT system in a short period of time". In *Proceedings of the Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities* (pp. 23-29): Second International Conference on Language Resources and Evaluation. Athens.

HUTCHINS, J. & SOMERS, H. (1992). *An introduction to Machine Translation*. New York: Academic Press Ltd.

WEINREICH, U. (ed.) (1968). Languages in contact. The Hague/Paris: Mouton.