THE AUTOMATIC TRANSLATION OF CLITIC PRONOUNS: THE ITS-2 SYSTEM*

Lorenza Russo (Lorenza.Russo@unige.ch)

1. INTRODUCTION

The interest in clitic pronouns in the generative grammar tradition goes back to the analysis made by Kayne (1975) as well as to the analysis on different types of clitics given by Zwicky (1977) (Miller & Monachesi, 2003: 67). Since then, clitics have been the object of much discussion in the linguistic literature and their definition is still a topic of debate. As pointed out by Kuchenbrandt et al. (2005: 1), various classification and tests for membership in a particular class of pronouns have been suggested in literature. In particular, traditional grammars distinguish two classes of pronouns, referred to as *strong* and *clitic*, in the sense that strong pronouns can receive emphatic stress and are morphosyntactically independent, while clitics depend on a host and tend to be unstressed.

According to Kayne (1975), we can say that in Romance languages clitic pronouns usually cliticize to a verbal host, while strong pronouns occur in nominal position. A clitic pronoun is then defined as an unstressed pronoun (1a), which cannot be conjoined (1b)[1] and which must be adjacent to the verb which constitutes its host (1c).

(1) a. * Je TE parle, pas LUI.
Je parle à TOI, pas à LUI.
'I speak to you, not to him.'

b. * Je **te** et **vous** parle.
Je parle à toi et à vous.
'I speak to you and you'.

c. Je **le** (* maintenant) dis.
Je le dis maintenant.
'I say it now.'

* The research described in this paper has been supported by a grant from the Swiss National Science Foundation, grant No 100015–130634. We would like to thank Lonneke van der Plas for a thorough reading of this paper.

[1] With respect to conjunction, Kayne (1975) notes not only that clitic pronouns cannot be conjoined (ia), but also that they cannot have wide scope on conjoined verbs (ib).

(i) a. * Jean **le** et **la** voit.
'Jean sees **him** and **her**.'

b. * Jean **le** voit et écoute.
'Jean sees **him** and listen.'

According to Miller & Monachesi (2003: 91), clitic pronouns cannot have wide scope except when the conjunction directly affects the verb which would be the clitic host in absence of conjunction. Consider the sentence in (Ib): if the verbs of that sentence would have been in the present perfect tense, the sentence "*Jean **l**'a vu et écouté*" ("Jean saw him and listened to him") would have been correct, because the clitic pronoun is adjacent to the auxiliary "*avoir*" and not to verbs "*voir*" and "*écouter*". This aspect is also true for Italian clitics. See Benincà & Cinque (1993) for more information.

The phenomenon of cliticization on a verbal host is crucial, not only for linguistic theory – for the comprehension for instance of the interactions between different parts of the grammar – but also for natural language processing (NLP) in general and machine translation in particular.

In this paper we discuss different types of problems that clitic pronouns can represent for a machine translation (MT) system, in order to highlight the necessity of automatic processing of this particular linguistic phenomenon. In particular, we focus on French clitic pronouns, for automatic translation from French to English. In some cases, we will refer to Italian clitics as well, from a comparative point of view, since they have a syntactic behavior quite similar to French clitics but also display a few differences. It is our aim to discuss difficult cases that often pose considerable difficulties in an MT system and which in our opinion need to be treated in order to produce both grammatically and syntactically correct translations.

Current MT systems[2] are often not able to recognize clitic pronouns in the source sentence. They generate, for instance, a target sentence in which no clitic pronoun appears (2a), confuse clitic pronoun gender and number (2b), generate it in an incorrect position in the target language (2c).[3]

(2) a. Je veux **les** attendre ici.
I want to wait here.[4]
'I want to wait **for them** here.'

b. Devo chieder**le** un documento.
I must ask them a document.[5]
'I must ask **her** a document.'

c. **Li** voglio comprare.
Je les veux acheter.[6]
'Je veux **les** acheter.'
'I want to buy **them**.'

This paper is organized as follows: the next section gives an overview of properties shared by strong pronouns and clitic pronouns which are problematic for natural language processing and automatic translation. Section 3 presents a number of other difficult cases, such as clitic clusters. In section 4, we describe the methodology applied by Its-2, an MT system developed in our laboratory, the LATL[7], in order to propose a possible approach to resolve problems in connection with clitic pronouns. Finally, section 5 provides conclusions and future work.

2. STRONG PRONOUNS, CLITIC PRONOUNS

As already mentioned, the clitics status in the pronominal system is a matter of debate. Starting in the mid–80s, in particular with Borer's (1984) and Suñer's (1993) works, clitic pronouns have been analyzed as agreement morphemes that are base-generated as part of a

[2] In this article, we will not consider the difference in performance between linguistic-based MT and a statistical MT for the current task of automatic translation of clitics. See Russo (2010) for an evaluation of the two different methodologies.
[3] See Russo (2011) on mistakes in the translation of clitic pronouns.
[4] Translation proposed by Google Translate (http://translate.google.com/#fr|en|).
[5] Translation proposed by Babelfish (http://fr.babelfish.yahoo.com/translate_txt).
[6] Translation proposed by Systran (http://www.systran.fr/).
[7] Language Technology Laboratory (http://www.latl.unige.ch).

clitic host. This alternative approach is different from the traditional one. It claims that strong pronouns in Romance languages behave like words, while clitic pronouns are expected to show affix-like properties (Kuchenbrandt et al., 2005: 6). This analysis has been adopted by many researchers and more recently by Miller & Monachesi (2003) who applied classical tests to the clitics in various Romance languages suggesting a parallel treatment of Romance object clitics and inflexional affixes. Kuchenbrandt et al. (2005) adopt a similar approach, suggesting that what has been traditionally referred to as clitic pronouns are in fact agreement markers.

In an opposite direction, researchers argued for a more fine-grained classification, primarily based on the morphosyntactic characteristics of pronouns. Laenzlinger (1993; 1998), for instance, suggests a further distinction between different types of *clitics* on the basis of their (weak vs strong) case features.[8] At the end of the 90s, Cardinaletti & Starke (1999) claim that the pronominal system can be divided into three major classes in terms of degrees of structural deficiency: *strong*, *weak* and *clitic* pronouns. Clitic pronouns are assumed to be deficient with respect to weak pronouns which are in turn deficient with respect to strong pronouns, distributionally, morphologically, semantically and prosodically (Cardinaletti & Starke, 1999: 216). In other words, strong pronouns are full nominal projections, while weak pronouns lack the highest functional layer and clitic pronouns lack both of the two highest functional layers (Cardinaletti & Starke, 1999: 214), that is, the CP layer and the ΣP layer. The latter is considered as the layer where prosodic features are realized (Cardinaletti & Starke, 1999: 195).[9]

In this article we are not going to treat weak pronouns in Cardinaletti & Starke's way, since our first aim is to focus on certain characteristics that clitic pronouns have in common with pronouns from an MT perspective, in order to underline the difficulties that these pronouns pose to an automatic system. This means that our analysis will combine the linguistic point of view, essentially syntactic, with the natural language processing point of view, trying to underline all the problems that arise from an automatic treatment of this linguistic phenomenon.

As clitic pronouns represent a type of personal pronouns, they share some characteristics with personal pronouns. These characteristics pose some problems for an MT system. Consider, for instance, the third-person singular personal pronouns in French. As shown in (3), to translate the French pronoun "*il*" to English, a human translator will choose

---

[8] An example of clitic and weak pronoun is given in (i) and taken from Laenzlinger (1998: 170).

(i) Donne-**le**-**moi**!
Donne-le[$_{clitic}$]-moi[$_{weak}$]
'Give **it to me**!'

Laenzlinger (1998) also calls incorporated and reduced clitic pronouns "syntactic clitics", while weak pronouns are defined as "LF-clitics", that is, clitics in Logical Form. See Laenzlinger (1998: 165-186) for more information.

[9] Here is an example, taken from Cardinaletti & Starke (1999: 212), of their theory of tripartition of clitic, weak and strong pronouns :

(i) a. Non **gli** dirò mai tutto. (clitic pronoun)
Not to-him I will tell never everything.
'I will never tell **him** everything.'
b. Non dirò mai **loro** tutto. (weak pronoun)
Not I will tell never to-them everything.
'I will never tell **them** everything.'
c. Non dirò mai tutto **a lui**. (strong pronoun)
Not I will tell never everything to-him.
'I will never tell **him** everything.'

between "*he*" and ''*it*", the first one referring to a human being, the second one to a thing. However, the machine translation system will select randomly one of the two personal pronouns.

(3) Il est beau.
'He / It is beautiful.'

Google Translate and Systran, in fact, translate the personal pronoun in the sentence in (3) as an impersonal pronoun (4). Of course, this is not the wrong translation, but it could be the wrong translation in a specific context.

(4) Il est beau.
It is beautiful.

If we consider the French clitic pronoun "*lui*" (5), it results in similar problem. It can be translated into two different pronouns in English, depending on its gender: "*him*" for the masculine, "*her*" for the feminine.

(5) Je **lui** donne un cadeau**.**
'I give **him** / **her** a present.'

In this case too, the human translator is able to solve the gender ambiguity in the context in which the pronoun appears, while an MT system is not able to apply the same analysis, usually preferring the masculine form instead of the feminine (6).

(6) Je **lui** donne un cadeau**.**
I give **him** a present.[10]

These cases of ambiguity result from the general problem of the pronominal references, which are a well-known problem for machine translation. Pronominal references occur not only in larger texts but also in simple sentences or in collocations. Consider, for instance, the text in (7).

(7) Sa nouvelle voiture est très puissante. Elle est contente de l'avoir achetée, d'autant plus qu'elle est très spacieuse.
'Her new car is very powerful. She's happy she bought it, especially since it is very roomy.'

In French, the third-person singular personal pronoun "*elle*" appears twice in the text: once, referring to the person who bought the car, and once referring to the car. Consider also that in this sentence the French possessive determiner "*sa*" is ambiguous as well for an MT system, because it can be translated in English as the masculine "*his*" or as the feminine "*her*", or even "*it*", according to the context. These types of pronominal references are very difficult to treat with an MT system. In fact, both Google Translate and Systran do not propose a good translation, the former proposing the masculine "*his*" for the possessive determiner, but a feminine subject (8i), while the latter proposes the impersonal form both for the possessive determiner as for the third-person singular personal pronoun (8ii).

[10] Translation proposed by Google Translate and by Systran.

(8) Sa nouvelle voiture est très puissante. Elle est contente de l'avoir achetée, d'autant plus qu'elle est très spacieuse.
(i) His new car is very powerful. She is happy to have purchased, especially since it is very spacious.
(ii) Its new car is very powerful. It is glad to have bought it, more especially as it is very roomy.

Consider now the example in (9).

(9) Ses quatre vérités, il **les lui** a dites dans un moment de rage.
His / her four truths, he these to-him / her told in a moment of anger.
'His / her few home truths, he tell **him** / **her** in a moment of anger.'

In this case, the pronominal reference appears in the idiomatic expression, "*dire ses quatre vérités à quelqu'un*" ("to tell him / her four truths to someone"), where the words of the expression are re-ordered and where the direct object "*ses quatre vérités*" is cliticized ("*les*") as well as the indirect object "*à quelqu'un*" ("*lui*").[11] As shown in (10i), this type of pronominal reference poses major problems for MT systems, which often encounter problem to translate the direct and the indirect cliticised objects. Moreover, MT systems often translate the multi-word expression in a word-by-word translation (10ii).

(10) Ses quatre vérités, il **les lui** a dites dans un moment de rage.
(i) Home truths, he said them in a moment of rage.[12]
(ii) Its four truths, he said them to him in one moment of rage.[13]

To better understand the problems that clitic pronouns pose for machine translation, note also that a French clitic pronoun can be confused, for instance, with a determiner in case of homography[14] (11a). Moreover, a clitic pronoun can be also combined with another clitic pronoun in a clitic cluster (11b), or in Italian it can move from the infinitive subordinate clause to the verb of the main sentence[15] (11c).

(11) a. Finalement il se décide et **l**'annonce à tout le monde.
* Finally it decides and the advertisement with everybody.[16]
'Finally he decides and announces **it** to everybody.'
b. Je **le lui** donne.[17]
I give him.[18]
'I give **it to him**.'

[11] See Nerima, Seretan & Wehrli (2006) and also Wehrli, Seretan & al. (2009) for a detailed description of collocation translation problems.
[12] Translation proposed by Google Translate.
[13] Translation proposed by Systran.
[14] On this aspect, Rizzi (2000: 96-97) suggest that in French the resemblance between accusative pronominal clitics ("*le*", "*la*", "*les*") and definite articles is not by chance. He explains this resemblance by the assumption that clitics are generated as special determiners somehow, which do not select any Noun Phrase.
[15] Consider that clitic climbing is a real problem for an MT system, for the translation from Italian, as shown in example (11c), as well as for the translation to Italian, because of the characteristic syntactic behavior of Italian clitics in such structures.
[16] Translation proposed by Systran. Note that Systran has problems also to translate the third-person singular personal pronoun "*il*".
[17] On clitic clusters see section 3.2 of this article.
[18] Translation proposed by Google Translate.

c. Voglio far**lo** ora. / **Lo** voglio fare ora.
I want it now. / I want to do now.[19]
'I want to do **it** now.'

Consider also that in the translation between two typologically different languages, as the case for French and English or for French and German, there will be other types of problems, above all because English and (standard) German do not have clitic pronouns strictly speaking. In this case, an MT system will have to transform the French clitic pronoun in a verb complement in English (12i) and in German (12ii).

(12) Je **lui** parle.
(i) I talk **to him**.
(ii) Ich spreche **mit ihm**.

3. DIFFICULT CASES

So far, we have considered quite simple sentences in which only one clitic pronoun appears, often an object (accusative or dative) clitic. To better understand the behavior of clitic pronouns[20] and to see how difficult they can be for automatic systems – especially for parsing and translation – consider some more difficult cases, such as the French clitic *se* and clitic clusters.

3.1. The French clitic *se*

According to Ruwet (1972) and to Wehrli (1986), among others, we can distinguish four cases of the French clitic *se*: reflexive / reciprocal *se* (13a); inherent *se* (13b); middle *se* (13c); and neuter *se* (13d).

(13) a. Marie **se** regarde dans le miroir.
Marie herself looks in the mirror.
'Marie looks at herself in the mirror.'
b. Marie **s**'évanouit.
Marie herself faints.
'Marie faints.'
c. Un veston de laine **se** lave facilement.[21]
A jacket of wool itself washes easily.
'A wool jacket washes easily.'
d. Le ciel **s**'éclaircit.
The sky itself brightens up.
'The sky brightens up.'

[19] Translations proposed by Google Translate.
[20] As far as the behavior of clitic pronouns is concerned, it is important to keep in mind that one of the first problem that clitics can pose to an MT system concerns their enclitic or the proclitic position. In this article, we do not discuss this problem, since it can be more or less difficult to treat on the basis of the language pair chosen and since we prefer to focalise our attention on more difficult cases concerning the French-English language pair.
[21] Example taken from Wehrli (1986: 266).

As a reflexive / reciprocal, clitic *se* can correspond to the direct object (14a) or the indirect object (14b); to the extra dative argument in the case of inalienable constructions (14c); or to an ethical dative (14d)[22].

(14) a. Marie **se** lave.
Marie herself washes.
'Marie washes **herself**.'
b. Marie et Jean **se** parlent souvent.
Marie and Jean to-each speak often.
'Marie and Jean often speak **to each other**.'
c. Marie **s**'est cassé la jambe.
Marie to-herself broke the leg.
'Marie broke her leg.'
d. Alors, on **se le** mange, ce melon?
'What are we waiting for to start eating that melon ?'

In this case, the difficulty for an MT system is to distinguish between these different uses of *se* since sometimes it is translated with the pronouns "*himself* / *herself*" [example in (14a), automatic translation in (15a)], sometimes it is not translated at all [as in examples in (13b-d; 14c); automatic translation in (15b-e)], or it is translated as "*to each other*" [example in (14b), automatic translation in (15f)].

(15) a. Marie **se** lave.
Marie washes. (Google Translate translation)
Marie washes herself. (Systran translation)
b. Marie **s**'évanouit.
Marie fainted. (Google Translate translation)
Marie disappears. (Systran translation)
c. Un veston de laine **se** lave facilement.
A wool jacket washes easily. (Google Translate translation)
A wool jacket is washed easily. (Systran translation)
d. Le ciel **s**'éclaircit.
The sky is clearing up. (Google Translate translation)
The sky is cleared up. (Systran translation)
e. Marie s'est cassé la jambe.
Mary broke her leg. (Google Translate translation)
Marie broke the leg. (Systran translation)
f. Marie et Jean **se** parlent souvent.
Mary and John talk frequently. (Google Translate translation)
Marie and Jean often speak themselves. (Systran translation)

As underlined in the examples in (15), one can observe that Google Translate encounters problem mostly with the reflective / reciprocal *se*, while Systran has problems with middle *se* and with the extra dative argument in the case of inalienable constructions.

[22] Example given by Wehrli (1986: 265) and taken from Morin (1981).

3.2. Clitic clusters

Let us now turn to clitic clusters. They occur when more than one pronoun cliticizes on the same verbal host. These types of combinations undergo some order restrictions that are not displayed by full arguments (16).

(16) Marie donne un livre à Jean. / Marie donne à Jean un livre.
Marie **le lui** donne. VS * Marie lui le donne.
'Marie gives **it to him**.'

The order of two clitic pronouns may also vary depending on the clitic pronouns found in the cluster. In French, usually the order is [Cl_{dat} – Cl_{acc}] with a first or second person clitic and a third person clitic[23] (17a-b) and [Cl_{acc} – Cl_{dat}] with two third person clitics (17c) (Laenzlinger, 1993).[24]

(17) a. Paul **te l**'achètera. (*Paul le t'achètera.)
Paul to-you it will buy.
'Paul will buy **it for you**.'
b. Paul **me la** dédie. (* Paul la me dédie.)
Paul to-me it dedicates.
'Paul dedicates **it to me**.'
c. Paul **le lui** dira demain. (* Paul lui le dira demain.)
Paul it to-him will say tomorrow.
'Paul will tell **him** tomorrow.'

But, for instance, as underlined by Laenzlinger (1998: 163), the proclitic order "*me le*" in French (18a) turns to the order "*le moi*" in enclisis (18b), the first person object pronoun is realized as a strong form in enclisis and occupies a final position in the cluster.

---

[23] The combination of an accusative first / second person clitic with a dative third person clitic is not permitted in French. To express this type of combination it is necessary to leave the dative pronoun in its full form.

(i) a. * Paul **me te** présente.
Paul me to-you introduces.
b. Paul **me** présente **à toi**.
'Paul introduces **me to you**.'
c. * Jean **me lui** présentera.
Jean will me to-him/to-her introduces.
d. Jean **me** présentera **à lui** / **à elle**.
'Jean will introduce **me to him** / **to her**.'

[24] In Italian too, the clitic cluster displays the order [Cl_{dat} – Cl_{acc}], but clitics are graphically combined as one word with the adjunction of the grapheme "e" (Cardinaletti & Starke, 1999: 169; Laenzlinger, 1998, among others), as shown in (V). See Cardinaletti (2007) for more information on "glielo" and on the other clitic clusters in Italian:

(i) a. gli_{dat} – lo_{acc}
Gli**e**lo
b. Paul **le lui** dira demain.
Jean **glielo** dirà domani.
'Jean will tell it to him tomorrow.'

(18) a. Marie **me le** montre.
Marie to-me it shows.
'Marie shows **it to me**.'
b. Montre-**le-moi**!
'Show **it to me** !'

The final position in the cluster can also be occupied by prepositional clitics. The French genitive / partitive clitic *en* and the oblique / locative clitic *y* can in fact be part of a clitic sequence as well, as shown in (19a-b), combining with direct / indirect object pronouns.

(19) a. Luka nous **en** parlera demain.
Luka to-us about it will talk tomorrow.
'Luka will talk **about it** tomorrow.'
b. Luka **vous y** a vu.
Luka you there saw.
'Luka saw you **there**.'

If we now consider clitic clusters from an MT system point of view, the fact that two clitic pronouns occur in the same sentence can affect the translation. Google Translate, for instance, translates only one clitic for sentences in (20a-d). However, Systran is able to correctly translate the clitic clusters, even though not the sentence in (20d).

(20) a. Marie **me le** montre.
Marie shows me. (Google Translate translation)
Marie shows it to me. (Systran translation)
b. Montre-**le-moi**!
Show me! (Google Translate translation)
Show it to me! (Systran translation)
c. Luka **vous y** a vu.
Luka you saw. (Google Translate translation)
Luka saw you there. (Systran translation)
d. Luka **nous en** parlera demain.
Luka we will talk tomorrow. (Google Translate translation)
Luka we will talk tomorrow. (Systran translation)

So far, we have highlighted most of the problems an MT system encounters concerning clitic pronoun translation. As showed in our examples, this specific linguistic phenomenon does not result in a correct translation very often. In the next sections we present the Its-2 translation system and propose a linguistically based approach to the automatic translation of clitic pronouns.

4. ITS-2

The Its-2 translation system, developed at the Language Technology Laboratory (Latl) in Geneva, relies on the deep syntactic parser Fips (Wehrli & Nerima, 2009). Fips is based on generative grammar concepts, some of which are specific to the language family supported by the system. It assigns a constituent structure of the type shown in (Figure 1) to the sentences analyzed (Wehrli, 2007) where XP is the maximal projection of the head X, while L and R

stands for (possibly empty) lists of the maximal projections corresponding, respectively, to the left and right subconstituents of X.[25]

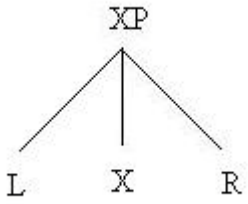

Figure 1: Fips X'- scheme

The Its-2 translation algorithm follows the traditional pattern of a transfer system with its three components:

1. lexical and syntactic analysis of the sentence in the source language;
2. lexical and grammatical transfer, from the source language to the target language;
3. syntactic and morphological generation of the sentence in the target language.

4.1. Methodology

The translation process starts with the lexical and syntactic analysis of the sentence in the source language, provided by Fips. This analysis determines the characteristics of lexical elements. The system produces an abstract sentence representation for the source language with an associated syntactic structure and a predicate-argument representation. Fips' parsing algorithm proceeds in a bottom-up fashion, by applying (general or language-specific) licensing rules, by treating alternatives in parallel and by using pruning heuristics. This so-called "right corner strategy" is driven by data and tries to attach a new element to a right-corner constituent in the left context. This left context is composed of a number of active nodes, to which the new element can be attached. The system takes all possible attachments into account in parallel. In other words, it attaches to a constituent a stack of active nodes[26], each of which constitutes a possible attachment on the right edge of the constituent.

Consider the following example:

(21) Jean lit le journal.
'Jean reads the newspaper.'

Fips identifies the source-language of the sentence in (21) and produces the structure in (22):

(22) $[_{TP}\ [_{DP}$ Jean] lit $[_{VP}\ [_{DP}$ le $[_{NP}$ journal]]]]

In the second step of the translation process, that is, the transfer from the source language to the target language, Its-2 goes through the phrase structure of the source language starting

[25] In this case, L and R are not equivalent to what generative linguists usually call Specifier and Complement. In our simplification, L and R are considered at the same level, while Specifier and Complement are not. X is a variable ranging over the set of lexical categories: Adv (adverb), A (adjective), N (noun), D (determiner), V (verb), P (preposition), C (conjunction), Interj (interjection), to which we add the morpheme T (Time/Inflexion) and the morpheme F (functional) for the secondary predication structures.

[26] We can define an active node stack as a stack where all possible constituents that can be attached as specifiers or as complements are positioned by the analyzer before being combined to the left context already specified.

from its head, and considers the right and left subconstituents. The lexical transfer occurs at the head-transfer level: the bilingual lexicon is used to map a source-language lexical item to an equivalent target-language item. Once the correspondence is found, the system projects the target-language structure on the basis of lexical information of the lexical item which is its head. To consider our example in (21), the system finds that *"lire"*, [$_{TP}$ [$_{DP}$] [$_{VP}$ lire]], is the head of the sentence and that it is translated in English as "*read*", [$_{TP}$ [$_{DP}$] [$_{VP}$ read]] according to the bilingual lexicon. On the basis of this information, Its-2 projects a new target-language structure. In other words, we assume that the lexical head determines a syntactic projection.

The transfer module is then used to pass from the source-language structure to the target-language structure. If this structure is similar to the source-language structure, no specific transfer rules are applied and the two structures are isomorphic.[27] The source-language structure and the target-language structure can also be very different, for instance, when the target-language object lexical item is not of the same case as the source-language object item.[28] The syntactic projection determines the target-language structure which lexical item is the head of.

In the final step, the target-language sentence generation is completed by the morphological generation of all words in the target language, as shown in (23). The morphological generation obtains the correct morphological form of every target-language words by considering all the local constraints such as number, gender, person, time and mode.

(23) [$_{TP}$ [$_{DP}$ Jean] [$_{VP}$ reads [$_{DP}$ the [$_{NP}$ newspaper]]]]

5. CLITIC PRONOUNS IN ITS-2

5.1. The analysis module

As far as the analysis and translation of clitic pronouns are concerned, the main idea for their processing is based on the generative grammar concept of clitic chain formation element associating an empty category in argument position to a clitic position.

The clitic analysis mechanism proceeds in two different steps: attachment and interpretation (Wehrli, 2007). As a clitic pronoun is read – whether as an independent word

[27] An example is given in (i), where the French syntactic construction and the Italian one are identical.

(i) a. Jean mange une pomme.
[$_{TP}$ [$_{DP}$ Jean] mange [$_{VP}$ [$_{DP}$ une [$_{NP}$ pomme]]]]
b. Gianni mangia una mela.
[$_{TP}$ [$_{DP}$ Gianni] mangia [$_{VP}$ [$_{DP}$ una [$_{NP}$ mela]]]]

[28] This phenomenon arises when constituents analyzed as arguments undergo a slightly different transfer process. In this case, the properties of the constituent in the target-language are in part determined by the subcategorisation features of the predicate in the target-language. Consider, for instance, the sentence in (iia) and its translation in English in (iib).

(ii) a. Paul regardait la voiture.
[$_{TP}$ [$_{DP}$ Paul] regardait [$_{VP}$ [$_{DP}$ la [$_{NP}$ voiture]]]]
b. Paul was looking at the car.
[$_{TP}$ [$_{DP}$ Paul] was [$_{VP}$ looking [$_{PP}$ at [$_{DP}$ the [$_{NP}$ car]]]]

The direct object of the French verb "*regarder*" ("look at") will be transferred in English as a prepositional phrase headed by the preposition "at", on the basis of the information in the bilingual lexicon. To be more precise, the bilingual lexicon specifies a correspondence between the French lexeme [$_{TP}$ [$_{DP}$] [$_{VP}$ regarder] [$_{DP}$]] and the English lexeme [$_{TP}$ [$_{DP}$] [$_{VP}$ look [$_{PP}$ at] [$_{DP}$]]]. For more details, see Russo & Wehrli (2011).

(24a) or as an orthographically attached an affix-like element (24b) – it is attached to the head of the verb form which follows it (enclisis) or which precedes it (proclisis).

(24) a. Je **le** lit.
'I read it.'
b. Dam**mi** il mio cappello!
Donne-moi mon chapeau!
'Give me my hat !'

Since the verbal host is not necessarily the one with respect to which the clitic pronoun can be interpreted – because it can be an auxiliary, for instance in (25) – a temporary data structure is used to store clitics until the parser has identified the main predicate of the sentence.

(25) Je **l'**ai fait volontiers.
'I have done **it** with pleasure.'

This temporary structure is also used to check the well-formedness of clitic sequences. Once the main predicate of the sentence is identified, the interpretation process starts. Clitic pronouns in the temporary data structure can then be interpreted as verb arguments (26a-b); as adjuncts of the verb (26c); as complements of a nominal argument (26d); or as arguments of a predicative adjective (26e).

(26) a. Je **lui** raconte un secret.
$[_{TP}\ [_{DP}$ Je] lui_i raconte $[_{VP}\ [_{PP}\ e_i\ ]\ [_{DP}$ un $[_{NP}$ secret]]]]
I tell **him** a secret.
$[_{TP}\ [_{DP}$ I] $[_{VP}$ tell $[_{DP}$ him] $[_{DP}$ a $[_{NP}$ secret]]]]
b. Paolo **en** parle ouvertement.
$[_{TP}\ [_{DP}$ Paolo] en_i parle $[_{VP}\ [_{AdvP}$ ouvertement $[_{PP}\ e_i\ ]\ ]\ ]\ ]$
Paolo speaks **about it** openly.
$[_{TP}\ [_{DP}$ Paolo] $[_{VP}$ speaks $[_{PP}$ about $[_{DP}$ it]] $[_{AdvP}$ openly]]]
c. Paolo **y** travaille souvent.
$[_{TP}\ [_{DP}$ Paolo] y travaille $[_{VP}\ [_{AdvP}$ souvent]]]
Paolo works **there** often.
$[_{TP}\ [_{DP}$ Paolo] $[_{VP}$ works $[_{AdvP}$ there] $[_{AdvP}$ often]]]
d. Paolo **en** connaissait la raison.
$[_{TP}\ [_{DP}$ Paolo] en_i connaissait $[_{VP}\ [_{DP}$ la $[_{NP}$ raison $[_{PP}\ e_i]\ ]\ ]\ ]\ ]$
Paolo knew the reason **for it**.
$[_{TP}\ [_{DP}$ Paolo] $[_{VP}$ knew $[_{DP}$ the $[_{NP}$ reason $[_{PP}$ for $[_{DP}$ it]]]]]]
e. Paolo **en** est fier.
$[_{TP}\ [_{DP}$ Paolo] en_i est $[_{VP}\ [_{FP}\ [_{DP}]\ [_{AP}$ fier $[_{PP}\ e_i]\ ]\ ]\ ]\ ]$
Paolo is proud **of it**.
$[_{TP}\ [_{DP}$ Paolo] is $[_{VP}\ [_{FP}\ [_{DP}]\ [_{AP}$ proud $[_{PP}$ of $[_{DP}$ it]]]]]]

5.2. The transfer and generation modules

Once the parse is completed, Its-2 starts the process of translation. We first need to find a translation for the main verb of the sentence in the bilingual lexicon for the lexical transfer. Thanks to the empty category associated to the clitic pronoun, the system can translate it properly. Consider for instance the sentence in (27).

(27) Je **l**'ai vu.
[$_{TP}$ [$_{DP}$ Je] l$_i$' ai [$_{VP}$ vu [$_{DP}$ e$_i$]]]
'I saw **him**.'

On the basis of the verb categorization, "*voir*" [$_{TP}$ [$_{DP}$] [$_{VP}$ voir] [$_{DP}$]], Its-2 analyses the French clitic "*le*" as the direct object of the verb "*voir*" (28a). If in the target language there are no clitic pronouns, then the direct objet of the sentence in (27) will be translate with a strong accusative pronoun (28b), as is the case in English.

(28) a. J'ai vu Cl(OD)[29].
[$_{TP}$ [$_{DP}$ J'] ai [$_{VP}$ vu [$_{DP}$ Cl]]]
b. J'ai vu Cl(OD).
I saw Pronoun(OD).
I saw him.

On the other hand, if the target language is a language that has clitic pronouns, as is the case for Italian, for instance, then the trace of the clitic pronoun in the source language (29a) is used to generate a target-language clitic pronoun in the object position (29b). Then, this pronoun is cliticised in the generation step, in order to obtain a sentence that is grammatically correct according to the target-language (29c).

(29) a. J'ai vu Cl(OD).
[$_{TP}$ [$_{DP}$ J'] ai [$_{VP}$ vu [$_{DP}$ Cl]]]
b. Ho visto Cl(OD).
[$_{TP}$ [$_{DP}$] ho [$_{VP}$ visto [$_{DP}$ Cl]]]
c. L'ho visto.
[$_{TP}$ [$_{DP}$] L$_i$' ho [$_{VP}$ visto [$_{DP}$ e$_i$]]]

Let us now consider translation from English to French, that is, from a source language that does not have clitic pronouns to a target language that does have clitic pronouns. Consider the sentence in (30):

(30) I will do **it**.

The English pronoun "*it*" (31a) is firstly translated as a direct object pronoun (31b) and then cliticized as a preverbal clitic in French (31c).

(31) a. I will do **it**.
[$_{TP}$ [$_{DP}$ I] will [$_{VP}$ do [$_{DP}$ it]]]
b. Je ferai Pronom(OD).[30]
c. Je **le** ferai.

This cliticization step – which holds for the translation from a clitic-language or from a non clitic language – is part of the generation module and involves the analysis of the possibility of cliticization of pronouns, as there are some constraints, for instance, on clitic clusters.

[29] Objet Direct (Direct object).
[30] In this case, also the pronoun "*je*" is considered as a phonological clitic, syntactically in a subject position.

5.2.1. *Translation of clitic clusters*

With respect to clitic clusters, consider the constraint on dative pronouns, which in French cannot be cliticized if they occur with a first and second person object clitic. Consider the case of the verb "*présenter quelqu'un à quelqu'un*" ("to introduce someone to someone"). As shown in (32), in French it is not correct to cliticize both pronouns.

(32) * Jean nous lui a présentés.
Jean nous a présentés à lui.
'Jean introduced us to him.'

This syntactic constraint is handled by the generation module and rules that are used in this module. The other way around, that is, for the translation from English to French, these rules are used as well to generate a correct sentence in the target language. Consider for instance the sentence in (33).

(33) Luka introduces **me to you**.

When the generation step starts, the rules used by this module do not allow the system to produce a sentence as given in (34).

(34) * Luka me te présente.

In particular, the system firstly considers the English direct object pronoun which is clitized as the clitic pronoun "*me*". Then, the indirect object "*to you*" is considered. Thanks to the rules that block the combination of first / second person accusative clitic to a third person dative clitic in French, Its-2 is able to produce the correct sentence, as shown in (35).[31]

(35) Luka **me** présente **à toi**.

[31] As far as clitic clusters is concerned, it would be more difficult to automatically translate clitic clusters in more specific contexts, such as causative constructions like "*faire* + infinitive". In this case, the combination of an accusative first or second person clitic and a dative third person clitic is allowed in French, but not one next to the other (i). As Perlmutter (1970) and Postal (1981) observed, the cluster "*me lui*" in French is only possible when a direct object clitic is in the middle of the cluster and only when the clitic pronoun "*me*" can be interpreted as the agent of the infinitive verb.

(i) Il **me le lui** a fait apporter.
'He let **me** give **it to him**.'

The "*me lui*" cluster is also accepted in a sentence like (ii) because the two pronouns do not function as arguments of the same predicate, as discussed in Laenzlinger (1998: 331): in fact, "*me*" is an argument of the verb "*semble*" ("seem"), while "*lui*" is an argument of the adjective "*infidèle*" ("unfaithful").

(ii) Elle **me lui** semble infidèle.
She to-me seems to-him unfaithful.
'She seems **to me** to be unfaithful **to him**.'

5.2.2. *Translation of French clitic se*

As far as the French clitic *se* is concerned, in our system we distinguish different realization of the clitic *se*, on the basis of the distinction discussed in section 3.1. When the system translates a clitic *se* with a function of direct / indirect object or a reflective / reciprocal *se*, the methodology applied is the one already discussed above. In particular, if there is no lexicalization in the mono- and bilingual lexicons, then Wehrli's (1986) assumptions are taken into account and the concept of absorption is applied. Wehrli (1986) proposes that the presence of the clitic *se* attached to a verb triggers a modification of the argument structure of this verb. Clitic *se* can absorbe any NP-type argument, that is, the subject argument – in case of middle *se* (36a), for instance; the direct or indirect object argument – in case of reflexive / reciprocal *se* (36b-c); the extradative argument – in case of inalienable possession constructions; the ethical dative.

(36) a. Ce livre **se vend** bien.
'This book sells well.'
b. Je **me** rase.
'I shave myself.'
c. **Nous nous** présentons nos amis.
We to-us introduce our friend.
'We introduce our friend **to each other**.'

This absorption modifies the argument structure of the verb which will have different properties depending on the argument that has been absorbed. Consider, for instance, the example in (37), taken from Wehrli (1986 : 271).

(37) Jean fait **se laver** les enfants.
Jean makes self-wash the kids.
'Jean makes the kids wash theirself.'

In this case, the affixation of clitic *se* to a transitive verb, such as "*laver*" ("to wash"), makes it syntactically intransitive.[32]

A different treatment is reserved for inherent *se* and to middle *se*. The inherent *se* can be found, for instance, with so-called pronominal verbs ("*se promener*" ("to walk"), "*s'évanouir*" ("to faint"), "*s'abstenir*" ("to abstain"))) as well as in some verbal expressions like "*s'en aller*" ("to go") or "*se la couler douce*" ("to take life easly"). The middle *se* case, can be found in verbs like "*se vendre*" ("to sell"). In these two cases, there is no analysis step for these inherent or middle *se* which are attached to the verb without being interpreted. In particular, for our examples given above, there is no analysis for the clitics "*s'en*" or "*se la*". These verbs are lexicalized in our mono- and bilingual lexicons, to be able to translate them correctly. Consider for instance the sentence (38).

(38) He went away.
'Il s'en est allé.'

In this case, in our bilingual lexicon "*s'en aller*" is the translation of "*to go away*". As there is no analysis and no transfer step, the verb "*s'en aller*" forces the cliticisation of the clitic "*se*"

[32] For more information on absorption, see Wehrli (1986 : 267-274).

and the clitic "*en*", which are attached as inherent clitics. The same process is applied when translating middle "*se*". Verbs like "*se vendre*" ("to sell") are then introduced in our mono- and bilingual lexicons. This solution simplifies the translation of this kind of clitics from French to English as well as from English to French.

6. CONCLUSION

In this paper, we focused on the characteristics of clitic pronouns that often pose considerable difficulties in an MT system. Our first aim was to highlight the necessity to automatically process this particular linguistic phenomenon in order to obtain grammatically correct automatic translations.

Secondly, we discussed a linguistic-based approach to parse and automatically translate clitic pronouns. Although there are some problems that are still left unsolved[33] – and we are still working on those – we have nevertheless made good progress on the cases discussed in this article. In particular, Its-2 can correctly translate clitic clusters as well as French clitic *se*, as discussed in the above section, from French to English and vice versa but also from French to other Romance languages, such as Italian.

For the latter, that is, the French-Italian translation and vice versa, we manually evaluated the translation of clitic pronouns in French and Italian, as discussed in other articles (Russo, 2010 and 2011). More specifically, we made a corpus of about 300 sentences for the translation from Italian to French and of about 300 sentences for the translation from French to Italian. We controlled the position of the clitic in the sentences as well as its case, gender and number and we tested three different syntactic structures: structures containing indicative tenses; infinitive structures; and structures containing a gerund. For this evaluation we tested three different systems: Its-2, Babelfish and Google Translate. The overall results for each system are summarized in the Table 1.

	FRENCH TO ITALIAN	ITALIAN TO FRENCH
ITS-2	88%	95%
Babelfish	61%	71%
Google Translate	18%	16%

Table 1. Percentage of correct translations of clitic pronouns

As shown, the overall percentage obtained by Its-2 confirms the fact that the methodology we built up for the automatic translation of clitic pronouns – as we discussed above – improves the translations, even if it is more difficult to develop and it takes more time to improve since it is based on a linguistic approach.

On the other hand, the low percentages obtained by Google Translate – as well as all the problems we highlighted in this article – confirm that it is necessary to improve the translation strategy of MT system in order to correctly translate this particular linguistic phenomenon we tested and discussed.

As far as Its-2 is concerned, we are currently improving rules for the translation of clitic clusters, as well as rules used in the analysis and in the generation modules. Moreover, we are conscious that some translation mistakes – such as the translation of a dative clitic pronoun in an accusative clitic pronoun (39a) or a missed cliticization of the clitic pronoun to its verbal host (39b) – is a direct result of the degree of improvement of the rules used by the system.

[33] That is the case, for instance, for the translation of clitic clusters such as the ones discussed in the note n. 31.

(39) a. Devo chieder**le** un documento.
Je dois **leur** demander un document.
'I have to ask **you** a document.
b. Il doit **me** répondre.
* Deve rispondere **mi**.
'He must answer **me**.'

7. REFERENCES

Benincà, P. & Cinque, G. (1993) "Su alcune differenze tra enclisi e proclisi", in *Omaggio a Gianfranco Folena*, Editoriale programma, vol. 3, 2314-2326.

Borer, H. (1984) *Parametric Syntax. Case studies in Semitic and Romance Languages.* Dordrecht, Foris.

Cardinaletti, A. (2007) "On different types of clitic clusters", in *Working Papers in Linguistics*, University of Venice, vol. 17, 27-76.

Cardinaletti, A. & Starke, M. (1999) "The typology of structural deficiency: A case study of the three classes of pronouns", in *Clitics in the Languages of Europe*, van Riemsdijk (editor), H., Berlin et New York, Mouton de Gruyter, 145-233.

Kayne, R. S. (1975) *French Syntax. The Transformational Cycle*. Cambridge, The MIT Press.

Kayne, R. S. (1991) "Romance Clitics, Verb Mouvement and PRO", in *Linguistic Inquiry*, vol. 22, n. 4, 647-686.

Kuchenbrandt, I., Kupish, T. and Rinke, E. (2005) "Pronominal Object in Romance. Comparing French, Italian, Portuguese, Romanian and Spanish", in *Working papers in Multilingualism*, vol. 67, Universität Hamburg.

Laenzlinger, C. (1993) "A syntactic view of Romance pronominal sequences", *Probus 5*, Walter de Gruyter, 241-270.

Laenzlinger, C. (1998) "Pronouns", in *Comparative Studies in Word Order Variation. Adverbs, Pronouns and Clause Structure in Romance and Germanic*, Amsterdam/Philadelphia, John Benjamins Publishing Company, 123-241.

Laenzlinger, C. (2003) *Initiation à la Syntaxe formelle du français*, Berne, Peter Lang.

Miller, P. & Monachesi, P. (2003) "Les pronoms clitiques dans les langues romanes", in *Les langues romanes: Problèmes de la phrase simple*, under the direction of Godard, D., Editions du CNRS, 67-123.

Morin, Y-C. (1981) "Some Myths About Pronominal Clitics in French", in *Linguistic Analysis*, 8.2, 95-109.

Nerima, L., Seretan, V. & Wehrli, E. (2006) "Le problème des collocations en TAL", in *Nouveaux cahiers de linguistique française*, n. 27, Genève, Suisse, 95–115.

Perlmutter, D. M. (1970) "Surface structure constraints in syntax", *Linguistic Inquiry*, 1, 187-255.

Postal, P. (1981) "The French cohesive infinitive construction", *Linguistic Analysis*, 8.3, 281-323.

Rizzi, L. (2000) "Some notes on Romance cliticization", in *Comparative syntax and language acquisition*, London and New York, Routledge leading linguists.

Russo, L. (2010) "La traduction automatique des pronoms clitiques. Quelle approche pour quels résultats?", in *Actes de la 17e Conférence sur le traitement automatique des langues naturelles (TALN 2010)*, Université de Montréal, Québec (Canada).

Russo, L. (2011 - in press) "La traduction automatique entre langues proches : les pronoms clitiques en italien et en français", in *Actes des XXIVes Journées de linguistique (JDL)*,

Université Laval (Québec), under the direction of the “Association des étudiants diplômés et inscrits en langues et linguistique” (A ´EDILLT) in collaboration with the “Centre interdisciplinaire de recherches sur les activités langagières” (CIRAL), Québec (Canada).

Russo, L. & Wehrli, E. (2011 – in press) "Traduction automatique et aide terminologique: le traducteur de mots en contexte TWiC et le traducteur de phrases Its–2", in *Traduttori e traduzioni*, Vallini, C., De Meo, A. & Caruso, V. (editors), Napoli, Liguori.

Ruwet, N. (1972) *Théorie syntaxique et syntaxe du français*, Le Seuil, Paris.

Suñer, M. (1993) "El papel de la concordancia en las construcciones de reduplicación de clíticos", in Fernández Soriano, O. (ed.) *Los pronombres átonos* (Gramática del español), Taurus Universitaria, 174-204

Wehrli, E. (1986) "On some properties of French clitic *se*", in *Syntax and Semantics*, H. Borer (editor), vol. 19, Academic Press Inc, 263-283.

Wehrli, E. (2007) "Fips, a "Deep" Linguistic Multilingual Parser", in *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*, Prague, Czech Republic, Association for Computational Linguistics, 120-127.

Wehrli, E. & Nerima, L. (2009) "L’analyseur syntaxique Fips", Journée thématique Atala *"Quel analyseur syntaxique pour le français?"*, Université Paris Diderot - Paris 7, Paris, France. Article available at: http://alpage.inria.fr/iwpt09/atala/fips.pdf

Wehrli, E., Seretan, V., Nerima, L., & Russo, L. (2009) "Collocations in a Rule-based MT System: A Case Study Evaluation of their Translation Adequacy", in *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, Barcelona, Spain, 128-135.

Zwicky, A. (1977) *On clitics*, in Bloomington, Indiana University Linguistics Club.