

La traduction automatique des séquences clitiques dans un traducteur à base de règles. *

Lorenza Russo, Éric Wehrli
Laboratoire d'Analyse et de Technologie du Langage (LATL)
Département de linguistique – Université de Genève
2, rue de Candolle – CH-1211 Genève 4
{Lorenza.Russo, Eric.Wehrli}@unige.ch

Résumé. Dans cet article, nous discutons la méthodologie utilisée par Its-2, un système de traduction à base de règles, pour la traduction des pronoms clitiques. En particulier, nous nous focalisons sur les séquences clitiques, pour la traduction automatique entre le français et l'anglais. Une évaluation basée sur un corpus de phrases construites montre le potentiel de notre approche pour des traductions de bonne qualité.

Abstract. In this paper we discuss the methodology applied by Its-2, a rule-based MT system, in order to translate clitic pronouns. In particular, we focus on French clitic clusters, for automatic translation between French and English. An evaluation based on a corpus of constructed sentences shows the potential of this approach for high-quality translation.

Mots-clés : Analyseur syntaxique, traduction automatique, pronom clitique, séquences clitiques.

Keywords: Syntactic parser, automatic translation, clitic pronoun, clitic clusters.

1 Introduction

Le phénomène de la cliticisation des pronoms a suscité l'attention de très nombreux linguistes, en particulier suite aux travaux de Kayne (1975). Mais les pronoms clitiques sont importants aussi du point de vue du traitement automatique du langage naturel (TALN) en général et de la traduction automatique en particulier. Les systèmes de traduction automatique actuellement disponibles ne sont pas toujours capables de reconnaître les pronoms clitiques dans la langue source¹. Ainsi, par exemple, dans le cas de séquences de pronoms clitiques, c'est-à-dire de l'occurrence de deux (ou plus) pronoms clitiques attachés au même hôte verbal, les systèmes de traduction automatique actuellement disponibles génèrent fréquemment des phrases cibles dans lesquelles seulement un des deux clitiques est traduit (1). Google Translate², par exemple, atteint un pourcentage très bas de traductions correctes des séquences clitiques sur un corpus d'exemples construits³.

Au vu de ces résultats et compte tenu aussi de l'absence de travaux de recherche à ce sujet, notre but est celui de souligner ici l'importance de l'information lexicale et syntaxique pour le traitement de telles constructions afin d'obtenir des traductions automatiques syntaxiquement correctes et complètes. Pour cela, nous présentons dans cet article la méthodologie utilisée par Its-2 – un traducteur automatique multilingue à base de règles développé dans notre laboratoire - en nous focalisant en particulier sur la paire de langues français-anglais et sur les séquences clitiques.

*. Le travail de recherche présenté ici a bénéficié du support du Fond National Suisse de la Recherche Scientifique (No 100015-130634). Cet article a été en partie adapté de l'article de Russo (2010).

1. Considérons, de plus, que dans la traduction de deux langues typologiquement différentes, comme c'est le cas pour le français et l'anglais ou aussi pour le français et l'allemand, par exemple, le problème principal est dû au fait que l'anglais et l'allemand standard n'ont pas de pronoms clitiques à proprement parler. Dans ce cas, un système de traduction automatique doit transformer le pronom clitique français (ia) dans un complément du verbe en anglais (ib) et en allemand (ic).

(i) a. Je lui parle. b. I talk to him. c. Ich spreche mit ihm.

2. En français Google traduction (<http://translate.google.com>).

3. Pour plus de détails sur cette évaluation et sur le type de corpus utilisé, nous renvoyons le lecteur à la section 3 de cet article.

- (1) Je le lui donne.
 I give him. (traduction proposée par Google Translate)
 I give it to him.

Cet article comprend trois parties. Suite à cette brève introduction, la section 2 décrit la stratégie de traduction de Its-2 appliquée aux séquences clitiques ; la section 3 est consacrée à l'évaluation des traductions proposées par notre système, ainsi que celles proposées par Google Translate et par Systran⁴, sur un corpus de phrases contenant des séquences clitiques. Finalement, la section 4 contient les conclusions de notre travail.

2 Its-2

Its-2 (Russo & Wehrli, 2011) est un système de traduction basé sur l'analyseur syntaxique profond Fips (Wehrli, 2007). Le processus de traduction se compose de trois phases principales : lors de la première phase, l'analyse lexicale et syntaxique de la phrase source, à l'aide de Fips, détermine la nature des éléments lexicaux et produit une représentation abstraite de la phrase source avec sa structure arborescente, ainsi qu'une représentation des relations entre prédicat (le verbe) et arguments (son sujet et ses compléments). Pendant la deuxième phase, le système traverse itérativement la structure arborescente source créée par Fips, repère les têtes lexicales et interroge le lexique bilingue pour trouver des correspondances de ces têtes lexicales dans la langue cible. Sur la base de la correspondance lexicale trouvée, ainsi que des informations syntaxiques contenues dans le lexique bilingue et dans le lexique monolingue cible, Its-2 projette une structure abstraite cible. Finalement, lors de la troisième phase, celle de la génération de la phrase cible, des transformations syntaxiques peuvent s'appliquer à la structure abstraite, par exemple pour déplacer des constituants, cliticiser des pronoms, déterminer les temps et modes verbaux, etc. Le processus de génération s'achève par la génération morphologique, qui sélectionne la forme morphologique appropriée pour chaque mot cible en tenant compte des valeurs morpho-syntaxiques telles que le nombre, le cas, le genre, la personne, le temps ou le mode, etc.

2.1 Traitement des pronoms clitiques en Its-2

Dans les lexiques mono et bilingues, l'information associée aux pronoms clitiques est relativement simple. Etant donné leur distribution particulière, nous assignons aux pronoms clitiques une catégorie lexicale spécifique, celle de clitique. En plus des valeurs morpho-syntaxiques de genre, nombres et personnes, les clitiques appartiennent à des sous-catégories distinctes, qui déterminent en partie leur interprétation mais surtout l'ordre dans lequel ils apparaissent au sein d'une chaîne. En français, nous distinguons 5 sous-catégories de clitiques (*se, le, lui, y, en*). Dans le dictionnaire bilingue français-anglais, les pronoms clitiques du français sont mis en correspondance avec des pronoms forts de l'anglais, par exemple *le* est en correspondance avec *it*.

Pour ce qui est de l'analyse et de la traduction des pronoms clitiques, par contre, l'idée principale pour leur traitement se base sur une notion empruntée à la grammaire générative, celle de la formation d'une chaîne clitique reliant un pronom clitique à une catégorie vide postverbale. Autrement dit, les pronoms clitiques – à l'exception des clitiques inhérents⁵ – sont associés à une catégorie vide en position d'argument ou d'ajout. Le mécanisme d'analyse des clitiques se fait en deux étapes : l'attachement et l'interprétation (Leoni de Léon & Michou, 2006; Wehrli, 2007). Quand le pronom clitique est lu – soit comme élément graphiquement séparé du verbe (2a) soit comme élément graphiquement attaché à celui-ci (2b) – on l'attache à la tête verbale qui le suit (proclise) (2a) ou qui le précède (enclise) (2b). La tête verbale correspond soit au verbe principal, dans le cas d'une phrase à temps simple (2a-b), soit au premier auxiliaire, dans le cas d'une phrase à temps composé (2c).

- (2) a. Tu le manges. b. Dimmi la verità ! c. Tu l'as cassé.
 You eat it. Dis-moi la vérité ! You have broken it.

Afin d'analyser au mieux les pronoms clitiques, on utilise une structure de données temporaire pour les stocker jusqu'à ce que l'analyseur syntaxique identifie le verbe principal de la phrase. Cette structure temporaire est aussi utilisée pour contrôler la bonne formation de la séquence clitique. Une fois que le verbe principal de la phrase est

4. Disponible à la page : <http://www.systran.fr>.

5. Un clitique inhérent ou lexical est un clitique qui ne correspond pas à un complément ou à un ajout du verbe hôte, comme par exemple le clitique *se* associé à un verbe essentiellement pronominal, comme *se suicider* ou *se lever*. À ce propos, voir Wehrli (1986) et Russo (2010).

identifié, le processus d'interprétation commence. Les pronoms clitiques dans la structure de données temporaire peuvent donc être interprétés comme arguments du verbe (3a) ; comme ajouts du verbe (3b) ; comme compléments d'un argument nominal (3c), ou encore comme arguments d'un adjectif prédicatif (3d)⁶.

- (3) a. Je lui donne un paquet.
 $[_{TP} [_{DP} \text{Je}] \text{ lui}_i \text{ donne } [_{VP} [_{DP} \text{un } [_{NP} \text{paquet}] [_{PP} e_i]]]]$
- b. Tu y vas souvent.
 $[_{TP} [_{DP} \text{TU}] y_i \text{ vas } [_{VP} [_{AdvP} \text{souvent}] [_{AdvP} e_i]]]$
- c. Elle en connaît la cause.
 $[_{TP} [_{DP} \text{Elle}] \text{ en}_i \text{ connaît } [_{VP} [_{DP} \text{la } [_{NP} \text{cause } [_{PP} e_i]]]]]$
- d. Nous en sommes heureux.
 $[_{TP} [_{DP} \text{Nous}] \text{ en}_i \text{ sommes } [_{VP} [_{AP} \text{heureux } [_{PP} e_i]]]]$

2.2 Les modules de transfert et génération

Une fois l'analyse syntaxique achevée, Its-2 commence le processus de traduction, notamment avec le transfert de la tête verbale en cherchant dans le lexique bilingue une correspondance de celle-ci dans la langue cible. Quant aux pronoms clitiques, ils ne sont pas traduits directement, mais exclusivement par le biais de la trace qui leur est associée. Considérons par exemple la phrase (4a).

- (4) a. Tu nous as parlé. b. Tu as parlé *PronomClitique(ObjetIndirect)*.
 $[_{TP} [_{DP} \text{TU}] \text{ nous}_i \text{ as } [_{VP} \text{parlé}] [_{PP} e_i]]$ $[_{TP} [_{DP} \text{TU}] \text{ as } [_{VP} \text{parlé}] [_{PP} \text{Pronom}]]$

Sur la base de la sous-catégorisation du verbe, le pronom clitique "nous" est analysé comme l'objet indirect du verbe "parler" (4b). Autrement dit, la trace associée au pronom clitique est traitée comme une sorte de pronom abstrait en position d'objet indirect. Au cours du processus de transfert, ce pronom abstrait est transféré dans la structure cible, la correspondance lexicale étant déterminée par le clitique antécédent de ce pronom abstrait. Si la langue cible est une langue à pronoms clitiques, il est possible que ce pronom subisse le processus de cliticisation. Cela est le cas, par exemple, pour la plupart des clitiques dans la traduction du français vers l'italien. À titre d'exemple, la trace du pronom clitique dans la structure source (4a) est utilisée pour générer un pronom en position d'objet dans la langue cible (5a). Ce pronom est ensuite cliticisé lors de la génération, afin d'obtenir une phrase cible grammaticalement correcte (5b).

- (5) a. Hai parlato *Pronom(OI)*. b. Ci hai parlato.
 $[_{TP} [_{DP}] \text{ Hai } [_{VP} \text{parlato}] [_{PP} \text{Pronom}]]$ $[_{TP} [_{DP}] \text{ Ci}_i \text{ hai } [_{VP} \text{parlato}] [_{PP} e_i]]$

Pour ce qui est de la traduction du français vers l'anglais, par contre, l'objet indirect de la phrase en (4a) sera traduit par le pronom fort *us*, complément de la préposition *to*, comme illustré en (6).

- (6) Tu as parlé *PronomClitique(OI)*.
 $[_{TP} [_{DP} \text{TU}] \text{ as } [_{VP} \text{parlé}] [_{PP} \text{Pronom}]]$
 Tu as parlé *Pronom(OI)*.
 You spoke to *Pronom(OI)*.
 You spoke to us.

Considérons maintenant la traduction de l'anglais vers le français, c'est-à-dire d'une langue sans pronoms clitiques vers une langue à clitiques. Pour la phrase (7a), le pronom anglais "it" est d'abord transféré dans la langue cible en tant que pronom objet direct (7b) et ensuite cliticisé (7c). Le processus de cliticisation fait partie du module de génération du français qui est soumis à des contraintes de placement et d'ordre comme celle, entre autres, sur les séquences clitiques, discutée dans la section suivante.

- (7) a. She will read it. b. Elle lira *Pronom(OD)*. c. Elle le lira.
 $[_{TP} [_{DP} \text{She}] \text{ will } [_{VP} \text{read } [_{DP} \text{it}]]]$

6. Dans ces deux derniers cas de figure (3c-d), le pronom clitique sera interprété quand le syntagme nominal ou l'adjectif de la phrase est identifié par le parseur.

2.3 Traduction des séquences clitiques

Comme déjà mentionné dans l'introduction, on peut définir une séquence clitique comme l'occurrence de plus d'un clitique associé au même hôte verbal. En ce qui concerne la traduction des séquences clitiques du français vers l'anglais et vice versa, les stratégies d'analyse et de transfert utilisées par Its-2 sont en grande partie similaires à celles déjà discutées dans les sections 2.1 et 2.2. En guise d'exemple, considérons l'exemple donné en (9a). Lors du transfert, le système considère d'abord la trace associée au pronom clitique à fonction d'objet direct ("le") et ensuite la trace associée au pronom clitique à fonction d'objet indirect ("te") (9b). La trace associée au pronom clitique objet direct est transférée comme un pronom objet direct en anglais, alors que la trace associée au pronom clitique objet indirect est transférée comme un pronom objet indirect (9c).

- (9) a. Il te le donnera.
 b. [_{TP} [_{DP} Il] te_j le_i donnera [_{VP} [_{DP} e_i] [_{PP} j]]]
 c. [_{TP} [_{DP} He] will [_{VP} give [_{DP} it] [_{PP} to [_{DP} you]]]]

Pour ce qui est du processus de génération, ce dernier devient légèrement plus compliqué quand il existe des contraintes syntaxiques sur les séquences clitiques. En français c'est le cas, par exemple, des contraintes sur les pronoms clitiques objet indirect, ces derniers ne pouvant pas être cliticisés s'ils se retrouvent avec un pronom clitique objet de première ou de deuxième personne (10a) (Perlmutter, 1971). Considérons, en particulier, le cas du verbe "*présenter quelqu'un à quelqu'un*". Comme montré en (10b), il n'est pas correct en français de cliticiser les deux pronoms.

- (10) a. * Mon père me lui promet.
 Mon père me promet à lui/à elle.
 b. * Jean nous lui a présentés.
 Jean nous a présentés à lui.

Cette contrainte syntaxique est présente dans l'analyseur Fips et permet de bloquer l'interprétation de phrases telles que celles données en (10) (et de les signaler comme agrammaticales). Dans l'autre sens de traduction, c'est-à-dire de l'anglais vers le français, les contraintes sur les séquences de pronoms clitiques sont intégrées au processus de cliticisation, dans le module de génération de la structure cible. En simplifiant quelque peu les choses, la contrainte dit simplement qu'un complément d'objet indirect ne peut pas être cliticisé en présence d'un clitique de la classe *se*⁷. C'est ce qui se passe dans la traduction de l'exemple (11a). Après cliticisation de l'objet direct, la présence du clitique *me* empêche la cliticisation du pronom objet indirect, qui donnerait la phrase agrammaticale (11b). Ce dernier reste donc dans sa position postverbale et nous obtenons la phrase grammaticale (11c)⁸.

- (11) a. Luka introduces me to you. b. * Luka me te présente. c. Luka me présente à toi.

3 L'évaluation

Afin d'évaluer Its-2 sur la traduction des séquences clitiques, nous avons rédigé un petit corpus d'exemples construits pour la traduction du français vers l'anglais. Chaque phrase dans le corpus présente une séquence de deux clitiques en position proclitique (préverbale) dans une structure syntaxique composée par sujet-séquence clitique-verbe. Les verbes sont conjugués au présent ou au futur. Dans la Table 1, on a schématisé les types de séquences clitiques présentes dans le corpus, ainsi que le nombre de phrases pour chaque séquence clitique. Comme on l'a mentionné dans l'introduction, nous avons testé sur le même corpus deux systèmes de traduction commercial différents du nôtre, Google Translate et Systran, afin de comparer les résultats obtenus. Comme montré dans la Table 2, Google Translate atteint un pourcentage total de traductions correctes des séquences clitiques de 15.3%.

7. Les règles de cliticisation sont ordonnées, la cliticisation de l'objet direct précédant celle de l'objet indirect.

8. Remarquons qu'il serait beaucoup plus difficile de traduire automatiquement les séquences clitiques dans des contextes plus spécifiques, comme par exemple les constructions causatives telles que "faire + infinitif". Dans ce cas, la combinaison d'un clitique objet direct de première/deuxième personne avec un clitique objet indirect de troisième personne est permis en français seulement si les deux clitiques sont séparés par un autre clitique (iia). Comme observé par Postal (1981), en français la séquence clitique "*me lui*" est possible seulement si un clitique objet direct est à l'intérieur de la séquence et seulement si le pronom clitique "*me*" peut être interprété comme l'agent du verbe à l'infinitif. La séquence "*me lui*" peut aussi être acceptée dans une phrase comme celle en (iib) parce que les deux pronoms ne fonctionnent pas comme arguments du même prédicat, comme souligné par Laenzlinger (1998) : en fait, "*me*" est l'argument du verbe "*sembler*" ("*seems*"), alors que "*lui*" est l'argument de l'adjectif "*infidèle*" ("*unfaithful*").

- (ii) a. Il me le lui a fait apporter.
 He let me give it to him.
 b. Elle me lui semble infidèle.
 She seems to me to be unfaithful to him.

verbe comme “rencontrer” en (15a) sélectionne de préférence un objet direct ayant le trait [+ humain], alors qu’un verbe comme “lire” in (15b) sélectionne de préférence un objet direct ayant le trait [- humain].¹⁰ Grâce à ces informations, on pourrait traduire en anglais le pronom clitique “le” par le masculin “him” dans le premier cas de figure, et par la forme neutre “it” dans le deuxième. Une telle information est déjà présente dans notre base de données lexicale, mais elle n’est pas encore exploitée.

4 Conclusion

Dans cet article, nous avons présenté les problèmes que les pronoms clitiques posent à un traducteur automatique. En particulier, nous nous sommes focalisés sur la paire de langues français–anglais, afin de traiter la traduction des séquences clitiques tout en exposant la stratégie utilisée par notre système pour ce phénomène linguistique spécifique. Une évaluation sur un petit corpus d’exemples construits montre que notre système traduit correctement dans plus de 90% des cas. Bien que les résultats obtenus doivent être confirmés par une évaluation sur un corpus de taille plus importante, ainsi que sur d’autres systèmes de traduction automatique, ce que cette évaluation suggère est qu’un traitement correcte de la cliticisation nécessite une description syntaxique très fine basée sur une information lexicale détaillée.

Références

- KAYNE R. S. (1975). *French Syntax. The Transformational Cycle*. Cambridge : MIT Press.
- LAENZLINGER C. (1998). Pronouns. In *Comparative Studies in Word Order Variation. Adverbs, Pronouns and Clause structure in Romance and Germanic*, chap. 3, pp. 123–241. Amsterdam - Philadelphia : John Benjamins Publishing Company.
- LAPPIN S. & LEASS, H.J. (1994). An algorithm for pronominal anaphora resolution. In *Computational Linguistics*, 20(4), pp. 535–561.
- LEONI DE LÉON J. A. & MICHOU A. (2006). Traitement des clitiques dans un environnement multilingue. In P. MERTENS, C. FAIRON, A. DISTER & P. WATRIN, Eds., *Verbum ex machina : Actes de la 13e conférence sur le traitement automatique des langues naturelles (TALN 2006)*, Cahiers du Cental 2.1, pp. 541–550, Louvain-la-Neuve, Belgique : UCL Presses Universitaires de Louvain.
- MITKOV R. & EVANS, R. & ORĂSAN, C (2002). A new, fully automatic version of Mitkov’s knowledge-poor pronoun resolution method. In *Proceedings of CICLing ’02*, Mexico City.
- PERLMUTTER D. M. (1971). *Deep and Surface Constraints in Syntax*. New York : Holt, Rinehart and Wilson.
- POSTAL P. M. (1981). The French Cohesive Infinitive Construction. In *Linguistic Analysis*, vol. 8.3, pp. 281–323.
- RUSSO L. (2010). The Automatic Translation of Clitic Pronouns : The Its-2 System. In T. IHSANE & C. LAENZLINGER, Eds., *GG@G - Generative Grammar in Geneva*, vol 6 (2010), pp. 201–220, Genève : Université de Genève.
- RUSSO L. (2011). La traduction automatique entre langues proches : les pronoms clitiques en italien et en français. In GENEVIÈVE BERNARD BARBEAU & CAROLINE GAGNÉ & GUILLAUME LEBLANC, Eds., *Actes des XXIVes Journées de linguistique*, pp. 141-153, Québec : CIRAL.
- RUSSO L. & WEHRLI É. (2011). Traduction automatique et aide terminologique : le traducteur de mots en contexte TWiC et le traducteur de phrases Its-2. In C. VALLINI, A. DE MEO & V. CARUSO, Eds., *Traduttori e traduzioni*, pp. 301–310, Napoli : Liguori.
- WEHRLI É. (1986). On some properties of French clitic se. In *Syntax and Semantics*, vol. 19, pp. 263-283, H. Borer (eds), Academic Press Inc.
- WEHRLI É. (2007). Fips, a "Deep" Linguistic Multilingual Parser. In *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*, pp. 120–127, Prague, Czech Republic : Association for Computational Linguistics.

10. Remarquons que les verbes “rencontrer” et “lire” pourraient aussi sélectionner des objets directs ayant respectivement le trait [-humain] (iia) et le trait [+ humain] (iib). On considère, cependant, ce deuxième type de sélection moins fréquent par rapport à la sélection discutée dans le texte à l’exemple (15).

(iii)

a. Ce problème, je le rencontre souvent.

b. Baudelaire, je le lis avec plaisir.