TALN 2010, Montréal, 19–23 juillet 2010

La traduction automatique des pronoms clitiques. Quelle approche pour quels résultats? *

Lorenza Russo
Laboratoire d'Analyse et de Technologie du Langage
Département de linguistique – Université de Genève
2, rue de Candolle – CH-1211 Genève 4
Lorenza.Russo@unige.ch

Résumé. Dans cet article, nous abordons la problématique de la traduction automatique des pronoms clitiques, en nous focalisant sur la traduction de l'italien vers le français et en comparant les résultats obtenus par trois systèmes : Its-2, développé au LATL (Laboratoire d'Analyse et de Technologie du Langage) et basé sur un analyseur syntaxique profond ; Babelfish, basé sur des règles linguistiques ; et Google Translate, caractérisé par une approche statistique.

Abstract. In this article, we discuss the problem of automatic translation of clitic pronouns, focalysing our attention on the translation from Italian to French and comparing the results obtained by three MT systems : Its-2, developed at LATL (Language Technology Laboratory) and based on a syntactic parser ; Babelfish, a rule-based system ; and Google Translate, caracterised by a statistical approach.

Mots-clés : Analyseur syntaxique, traduction automatique, pronoms clitiques, proclise, enclise.

Keywords: Syntactic parser, automatic translation, clitic pronouns, proclisis, enclisis.

1 Introduction

Le système pronominal, en italien tout comme en français, est constitué de pronoms dits forts et de pronoms dits faibles, atones ou clitiques. Parmi les pronoms faibles, on définit un pronom clitique comme un pronom non accentué qui n'occupe pas la même place syntaxique qu'un pronom fort (1a), qui est conjoint au verbe qui lui sert d'hôte (1b) et qui ne peut pas subir de phénomènes de coordination (1c) (Kayne, 1975).

(1)	a.	Io voglio quella.	b.	Io **lo** (*adesso) dico.	c.	* Io **ti** e **vi** parlo.
		Je veux celle-là.		Je **le** (*maintenant) dis.		* Je **te** et **vous** parle.
		*Je **la** veux.*		*Je **le** dis maintenant.*		*Je parle à toi et à vous.*

∗. Nous remercions le Fonds National Suisse de la Recherche Scientifique qui a soutenu ce travail de recherche (No 100015-130634) ainsi que le directeur du projet décrit dans cet article, le Prof. Éric Wehrli, et Sandra Schwab pour une lecture attentive de ce texte.

Dans les deux langues qui font l'objet de ce travail – l'italien et le français –, les pronoms clitiques présentent des réalisations morphologiques différentes selon le cas qu'ils portent (accusatif ou datif) (TAB. 1) ainsi que des réalisations syntaxiques différentes par rapport à leur mode d'attachement. Dans ce dernier cas, on parle de position proclitique (préverbale) ou de position enclitique (postverbale) du pronom clitique par rapport au verbe auquel il s'attache.

	Clitiques accusatifs						**Clitiques datifs**					
Italien	*mi*	*ti*	*lo / la*	*ci*	*vi*	*li / le*	*mi*	*ti*	*gli / le*	*ci*	*vi*	*loro*
Français	*me*	*te*	*le / la*	*nous*	*vous*	*les*	*me*	*te*	*lui*	*nous*	*vous*	*leur*

TABLE 1 – Réalisations morphologiques des pronoms clitiques en italien et en français

Ainsi, dans le cas des structures syntaxiques à l'indicatif, l'italien et le français présentent le clitique en position proclitique (2a) [1], alors que dans les phrases présentant un gérondif l'italien préfère l'enclise et le français la proclise (2b). De plus, dans les structures infinitives subordonnées, l'italien présente l'enclise (2c) tout comme la proclise (2d) là où le français n'accepte que la proclise.

(2) a. **Lo** guardo.
*Je **le** regarde.*

b. Mi addormento leggendo**ti** una favola.
*Je m'endors en **te** lisant une fable.*

c. Voglio mangiar**la**.
*Je veux **la** manger.*

d. **La** voglio mangiare.
*Je veux **la** manger.*

Considéré du point de vue de la traduction automatique, le phénomène des pronoms clitiques se révèle difficile à traiter parce qu'un traducteur automatique risque, par exemple, de confondre le pronom clitique avec un déterminant en cas d'homographie (3a) ; de le générer dans une position d'attachement non adéquate dans la langue cible (3b) ; ou encore, de mal identifier son cas ou son genre (3c).

(3) a. **L**'annuncio ad alta voce.
L'annonce à haute voix.
*Je **l**'annonce à haute voix.*

b. **Le** inizio a preparare.
Je **les** commence à préparer.
*Je commence à **les** préparer.*

c. Paolo **gli** parla.
Paolo **en** parle.
*Paolo **lui** parle.*

Tout en gardant à l'esprit notre cadre de référence, à savoir la traduction des pronoms clitiques de l'italien vers le français, nous nous sommes posé les questions suivantes : est-ce que les deux principales approches en traduction automatique – linguistique et statistique – mènent à des différences dans la qualité des traductions proposées ? En cas de réponse positive, est-ce que l'une des deux approches se révèle plus efficace ? Afin de répondre à ces questions, nous avons testé sur un même corpus italien trois systèmes à notre avis représentatifs des deux principales approches en traduction automatique : Its-2 – sur lequel nous travaillons au LATL – basé sur un analyseur syntaxique profond ; Babelfish, élaboré par Systran et basé lui aussi sur des règles linguistiques ou expertes ; et Google Translate [2], basé sur une approche statistique.

1. Remarquons cependant que dans les phrases à l'impératif positif le clitique est enclitique en italien tout comme en français (*Di**gli** la verità ! – Dis-**lui** la vérité !*). Dans les phrases à l'impératif négatif, par contre, l'italien accepte les deux positions proclitique et enclitique alors que le français ne garde que la proclise (*Non **lo** vendere ! / Non vender**lo** ! – Ne **le** vends pas !*).

2. En français Google traduction (`http://translate.google.fr/#`).

2 Description des systèmes et du corpus

En ce qui concerne Its-2, il s'agit d'un traducteur automatique basé sur une représentation des constituants de la phrase ainsi que des relations entre prédicats et arguments. La stratégie de traduction de Its-2 se compose de trois phases principales : 1) l'analyse lexicale et syntaxique de la phrase source ; 2) le transfert lexical et syntaxique ; et 3) la génération morphologique et syntaxique de la phrase cible. Dans la première phase, l'identification de la nature des éléments lexicaux et grammaticaux de la langue source se fait à l'aide de l'analyseur syntaxique Fips (Wehrli & Nerima, 2009), inspiré des théories générativistes chomskyennes, qui produit une structure arborescente de la phrase source. Après avoir repéré la tête lexicale de la structure arborescente créée par Fips, Its-2 interroge, dans la deuxième phase, le lexique bilingue pour trouver une correspondance de cette tête lexicale 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 nouvelle structure cible, en considérant les autres éléments de la phrase, à savoir les éventuels sous-arbres gauche et droit. Finalement, lors de la troisième phase, il génère la bonne forme morphologique de chaque mot cible en tenant compte des contraintes locales telles que le nombre, le cas, le genre, la personne, le temps ou le mode. [3]

Pour ce qui est des deux autres systèmes qui font l'objet de notre évaluation, il est nécessaire de souligner que, puisque il s'agit de deux systèmes commercialisés, les informations disponibles relatives à leurs stratégies de traduction sont très peu nombreuses. Le premier système, Babelfish – une version de Systran (`http://www.systran.fr`) – fait partie des systèmes de traduction automatique à transfert qui bénéficient de règles linguistiques. Malheureusement, les informations disponibles concernant Systran ne contiennent pas de détails spécifiques concernant l'approche linguistique et les règles que le système utilise. Quelques informations d'ordre général sont disponibles à la page `http://wwwv5.systransoft.com/IDC/26459.html` et dans Yang & Lange (2003). Quant au deuxième système, Google Translate, il est basée sur une méthode statistique : des milliards de mots provenant de corpus monolingues ou de corpus alignés – créés à partir de traductions réalisées par des traducteurs professionnels – sont introduits dans le système. Des techniques d'apprentissage statistique sont ensuite appliquées pour créer le modèle de traduction. [4]

Le corpus utilisé pour notre évaluation ne compte que des phrases en italien, afin de tester la traduction de l'italien vers le français. Il a été rédigé manuellement afin d'être en mesure de contrôler certains aspects spécifiques (Lehmann & Oepen, 1996), tels que la structure syntaxique, le cas du clitique et sa position dans la phrase, afin d'éviter l'interaction entre différents phénomènes linguistiques, vu la complexité des phrases contenues dans des corpus plus développés. Nous avons donc pris en considération les structures syntaxiques que nous avons présentées dans la section 1 : les structures transitives non réfléchies à l'indicatif (4a) ; les phrases présentant un gérondif (4b) ; les structures infinitives subordonnées à proclise (4c) ; et les structures infinitives subordonnées à enclise (4d). Plus précisement, notre corpus se compose de cinq phrases pour chaque pronom clitique [5] et pour chaque structure syntaxique considérés (TAB. 2) pour un total de 295 phrases [6].

3. Compte tenu de l'ampleur du sujet ici traité et de l'espace limité à notre disposition, nous invitons le lecteur à consulter Russo & Wehrli (2010) pour plus d'informations sur la stratégie de traduction de Its-2 et Leoni de Léon & Michou (2006) pour une description détaillée du traitement automatique des pronoms clitiques dans Fips.

4. Consulter à ce propos `http://www.google.com/intl/fr/help/faq_translation.html`.

5. La complexité du phénomène traité nous a obligé d'exclure de notre corpus les combinaisons de clitiques (*glielo* / *le lui*), ainsi que les clitiques locatifs (*ci* / *y*), réflexifs (*se* / *se*) et partitifs (*ne* / *en*).

6. Comme indiqué en (TAB. 2), les structures infinitives à proclise comptent cinq phrases en moins parce que le clitique

(4) a. **Gli** invierò una cartolina.
*Je **lui** enverrai une carte postale.*

b. Sono felice guardando**ti** ballare.
*Je suis heureux en **te** regardant danser.*

c. **Lo** potete fare domani.
*Vous pouvez **le** faire demain.*

d. Maria vuole fare **loro** un disegno.
*Maria veut **leur** faire un dessin.*

Structures	**Clitiques accusatifs**								**Clitiques datifs**							**Total**
	mi	*ti*	*lo*	*la*	*ci*	*vi*	*li*	*le*	*mi*	*ti*	*gli*	*le*	*ci*	*vi*	*loro*	
transitives	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	75
gérondives	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	75
inf. procl.	5	5	5	5	5	5	5	5	5	5	5	5	5	5	0	70
inf. encl.	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	75
Total	20	20	20	20	20	20	20	20	20	20	20	20	20	20	15	295

TABLE 2 – Nombre de phrases pour chaque clitique et pour chaque structure syntaxique considérés

3 Résultats et discussion

Étant donné la taille et la nature du corpus testé, les résultats que nous présentons dans cette section ne peuvent être considérés que comme exploratoires. De plus, les pourcentages de traductions correctes donnés ici sont le résultat d'une évaluation manuelle, motivée par la taille du corpus ainsi que par la nécessité de nous concentrer sur la traduction du phénomène des pronoms clitiques et non pas sur la traduction de la phrase dans sa totalité. Nous avons considéré une traduction comme correcte seulement dans le cas où le clitique était présent dans la phrase cible, généré correctement dans son cas, genre et nombre ainsi qu'attaché à son réel nœud d'attachement dans la langue cible. Comme le montre le tableau ci-dessous (TAB. 3), Its-2 présente un pourcentage élevé de traductions correctes, soit des traductions bien meilleures que Babelfish et Google Translate, que cela soit pour les clitiques accusatifs ou datifs. Pour ce qui est de Babelfish, la traduction des pronoms accusatifs présente un pourcentage plus élevé que celle des clitiques datifs. Quant à Google Translate, il présente un faible taux de traductions correctes que cela soit pour les clitiques accusatifs ou datifs.

Clitiques	**accusatifs**	**datifs**
Its-2	99.37%	90.71%
Babelfish	81.87%	59.28%
Google Translate	17.5%	15.71%

TABLE 3 – Pourcentage de traductions correctes des pronoms clitiques

Si on examine les détails des résultats obtenus pour chaque clitique pris en considération (TAB. 4 et TAB. 5), Its-2 présente un pourcentage élevé de traductions correctes, avec toutefois un taux plus bas pour les pronoms clitiques datifs *gli* (85%), *le* et *vi* (80%) (TAB. 5). De même, Babelfish présente davantage de problèmes pour la traduction des pronoms clitiques datifs *le* (40%) et *loro* (0%), ce dernier n'étant jamais traduit (TAB. 5).

En ce qui concerne Google Translate, il présente un pourcentage très faible de traductions correctes des

loro (*leur*) n'est pas accepté dans ce type de constructions.

Clitiques accusatifs	*mi*	*ti*	*lo*	*la*	*ci*	*vi*	*li*	*le*
Its-2	100%	100%	100%	100%	100%	95%	100%	100%
Babelfish	95%	95%	80%	80%	80%	90%	70%	65%
Google Translate	20%	0%	40%	15%	25%	25%	10%	25%

TABLE 4 – Pourcentage de traductions correctes pour chaque clitique accusatif

pronoms clitiques accusatifs *la* (15%), *li* (10%) et *ti* (0%) (TAB. 4) ainsi que des pronoms clitiques datifs *mi* et *ci* (10%), *gli* et *le* (5%), et *ti* (0%) [7] (TAB. 5).

Clitiques datifs	*mi*	*ti*	*gli*	*le*	*ci*	*vi*	*loro*
Its-2	100%	100%	85%	80%	100%	80%	100%
Babelfish	70%	75%	75%	40%	75%	80%	0%
Google Translate	10%	0%	5%	5%	10%	35%	50%

TABLE 5 – Pourcentage de traductions correctes pour chaque clitique datif

Ces pourcentages se répercutent aussi sur les résultats considérés du point de vue des structures syntaxiques prises en compte dans notre corpus (TAB. 6 et TAB. 7). En général, les diverses structures ne posent pas de problèmes à Its-2 qui dans la plupart des cas propose une traduction correcte. Cependant, un pourcentage relativement plus bas est atteint pour la traduction des clitiques datifs dans les phrases avec un gérondif (88.57%) et dans les structures infinitives à enclise (74.28%) (TAB. 7). Babelfish, quant à lui, présente des pourcentages généralement élevés pour la traduction des pronoms clitiques accusatifs et datifs, à l'exception des structures infinitives à proclise, en particulier pour les clitiques datifs (8.57%) (TAB. 7).

Structures	**transitives**	**gérondives**	**inf. à proclise**	**inf. à enclise**
Its-2	100%	100%	100%	97.5%
Babelfish	90%	87.5%	62.5%	87.5%
Google Translate	25%	10%	15%	30%

TABLE 6 – Pourcentage de traductions correctes des clitiques accusatifs pour chaque structure syntaxique

En ce qui concerne les résultats obtenus par Google Translate, ce système n'atteint jamais plus de 40% de traductions correctes, que cela soit pour les pronoms clitiques accusatifs ou datifs. En particulier, le pourcentage de traductions correctes ne dépasse pas 10% pour les clitiques accusatifs dans les phrases présentant un gérondif (TAB. 6) et descend à 0% pour la traduction des pronoms clitiques dans les structures infinitives à proclise (TAB. 7).

Structures	**transitives**	**gérondives**	**inf. à proclise**	**inf. à enclise**
Its-2	100%	88.57%	100%	74.28%
Babelfish	77.14%	74.28%	8.57%	77.14%
Google Translate	40%	11.42%	0%	14.28%

TABLE 7 – Pourcentage de traductions correctes des clitiques datifs pour chaque structure syntaxique

7. La raison pour laquelle la traduction de *ti* (*te*) – accusatif et datif – est à 0% est que Google Translate le traduit dans la plupart des cas par la forme de politesse *vous*.

Pour revenir aux deux questions que nous nous sommes posées au début de notre travail, les résultats présentés ici nous permettent d'affirmer qu'une différence dans l'approche en traduction automatique mène aussi à une différence dans la qualité des traductions proposées. De plus, les pourcentages très élevés atteints par Its-2, ainsi que ceux relativement élevés de Babelfish, nous mènent à considérer l'approche linguistique comme plus fructueuse par rapport à une approche statistique comme celle de Google Translate, du moins pour le genre de corpus utilisé, composé de phrases simples et courtes, et pour la paire de langues choisie.

4 Conclusion

Dans cet article, nous avons abordé la problématique de la traduction automatique des pronoms clitiques de l'italien vers le français, et ce, en nous focalisant en particulier sur la traduction des pronoms clitiques dans différentes structures syntaxiques et en examinant l'influence sur les résultats de l'approche – linguistique ou statistique – utilisée. Les résultats ont montré une nette différence dans la qualité des traductions des systèmes étudiés : en particulier, un traducteur automatique bénéficiant de règles linguistiques atteint de meilleurs taux de traductions correctes qu'un système à base statistique. Si une telle conclusion, basée sur une étude exploratoire, est difficile à généraliser vu la taille du corpus utilisé ainsi que la nature de l'évaluation menée, et se doit d'être confirmée au moyen d'un corpus plus large, sur d'autres phénomènes linguistiques et sur d'autres paires de langues, elle montre néanmoins la validité du recours à des règles linguistiques notamment pour la traduction de phénomènes linguistiques très pointus et complexes comme celui qui a fait l'objet de cette evaluation.

Références

KAYNE R. S. (1975). *French Syntax. The Transformational Cycle*. Cambridge : MIT Press.

LEHMANN S. & OEPEN S. (1996). "TSNLP-Test Suites for Natural Language Processing." dans *Proceedings of the 16th Conference on Computational linguistics*, volume 2, p. 711 – 716, Copenhagen : Association for Computational Linguistics.

LEONI DE LÉON J. A. & MICHOU A. (2006). "Traitement des clitiques dans un environnement multilingue." dans 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, p. 541–550, Louvain-la-Neuve, Belgique : UCL Presses Universitaires de Louvain.

RUSSO L. & WEHRLI É. (2010). "Traduction automatique et aide terminologique : le traducteur de mots en contexte TWiC et le traducteur de phrases Its-2." dans C. VALLINI, A. DE MEO & V. CARUSO, Eds., *Traduttori e traduzioni*, Naples : Liguori.

WEHRLI É. & NERIMA L. (2009). "L'analyseur syntaxique Fips." dans *Actes of the 11th International Conference on Parsing Technologies (IWPT '09), sponsored by ACL/SIGPARSE* : Université Paris Diderot - Paris 7.

YANG J. & LANGE E. (2003). "Going live on the Internet." dans *Computers and Translation. A translator's guide*, p. 191–210. John Benjamins Publishing Company.