TALN 2011, Montpellier, 27 juin – 1er juillet 2011

La traduction automatique des pronoms. Problèmes et perspectives. *

Yves Scherrer, Lorenza Russo, Jean-Philippe Goldman,
Sharid Loáiciga, Luka Nerima, Éric Wehrli

Laboratoire d'Analyse et de Technologie du Langage
Département de Linguistique – Université de Genève
2, rue de Candolle – CH-1211 Genève 4

{yves.scherrer, lorenza.russo, jean-philippe.goldman,
sharid.loaiciga, luka.nerima, eric.wehrli}@unige.ch

Résumé. Dans cette étude, notre système de traduction automatique, Its-2, a fait l'objet d'une évaluation manuelle de la traduction des pronoms pour cinq paires de langues et sur deux corpus : un corpus littéraire et un corpus de communiqués de presse. Les résultats montrent que les pourcentages d'erreurs peuvent atteindre 60% selon la paire de langues et le corpus. Nous discutons ainsi deux pistes de recherche pour l'amélioration des performances de Its-2 : la résolution des ambiguïtés d'analyse et la résolution des anaphores pronominales.

Abstract. In this work, we present the results of a manual evaluation of our machine translation system, Its-2, on the task of pronoun translation for five language pairs and in two corpora : a litterary corpus and a corpus of press releases. The results show that the error rates reach 60% depending on the language pair and the corpus. Then we discuss two proposals for improving the performances of Its-2 : resolution of source language ambiguities and resolution of pronominal anaphora.

Mots-clés : Pronoms, traduction automatique, analyse syntaxique, anaphores pronominales.

Keywords: Pronouns, machine translation, parsing, pronominal anaphora.

1 Introduction

Il y a un quart de siècle, Kay (1986) affirmait : « We know a good deal more about programming techniques and have larger machines to work with ; we have more elegant theories of syntax and what modern linguists are pleased to call semantics ; and there has been some exploratory work on anaphora. But, we still have little idea how to translate into a closely related language like French or German, English sentences containing such words as *he*, *she*, *it*, *not*, *and*, and *of*. » La recherche que nous présentons dans cet article prend comme point de départ l'affirmation de M. Kay et se donne comme objectif l'évaluation de la traduction automatique des pronoms. La plupart des recherches récentes portant sur la résolution des anaphores – pas forcément dans une perspective de traduction (Mitkov *et al.*, 2007) –, nous avons lancé une évaluation plus vaste concernant plusieurs paires de langues. Notre objectif principal est d'améliorer notre système de traduction automatique sur ce phénomène : nous en avons donc repéré les principaux problèmes ainsi que des pistes de recherche pour les résoudre.

Dans la section 2, nous présentons rapidement une pré-étude sur la distribution des pronoms dans la langue source, pour ensuite donner, dans la section 3, les détails de l'étude des pronoms dans la langue cible. La section 4 présente les résultats obtenus ; enfin, dans la section 5, nous discutons les pistes de recherche pour de futures implémentations.

∗. Le travail de recherche ici présenté a bénéficié du support du Fonds National Suisse de la Recherche Scientifique (No 100015-130634).

2 Les pronoms dans la langue source

Une partie du problème de la traduction des pronoms dépend de la bonne identification des formes pronominales dans la langue source. Pour ce qui concerne le français, Tutin (2002) et Laurent (2001), entre autres, ont mené des études sur corpus pour déterminer la distribution des pronoms dans différents genres de textes. Dans une optique similaire, mais orientée vers la traduction automatique et donc plurilingue, nous avons fait une pré-étude pour mieux déterminer les distributions des pronoms dans la langue source (Russo *et al.*, 2011). Nous voulions quantifier l'influence du style de texte, de la langue et du type de pronom. Dans ce but, nous avons annoté à l'aide de l'analyseur syntaxique Fips (Wehrli, 2007) deux corpus différents : un corpus littéraire, le *Petit Prince*, en français, anglais, allemand et italien (environ 1600 phrases) [1] et un corpus de communiqués de presse de l'Administration Fédérale Suisse, disponible dans les mêmes langues (environ 1000 phrases) [2]. Nous avons repéré tous les mots qui ont été étiquetés par Fips comme pronoms personnels (*je, eux, nous*) [3], clitiques (*y, la*), démonstratifs (*ça, cela*), relatifs (*qui, que*), indéfinis (*chacun, personne*), interrogatifs (*qui*) et déterminants possessifs (*ma, leurs*).

	Petit Prince				**Presse**			
Langue	Mots	Pronoms	Dont PERS	Dont CLI	Mots	Pronoms	Dont PERS	Dont CLI
FR	14 864	20,1%	47,5%	21,8%	26 995	3,1%	23,9%	24,1%
EN	16 495	17,9%	71,3%		23 079	2,4%	38,8%	
IT	13 197	11,0%	15,8%	48,7%	26 286	2,2%	5,4%	40,2%
DE	14 007	19,2%	69,9%		20 882	2,9%	37,2%	

TABLE 1 – Distribution des pronoms selon les langues et les corpus.

La Table 1 montre les pourcentages d'occurrence des pronoms. On constate d'une part que notre corpus littéraire contient sensiblement plus de pronoms que notre corpus journalistique, principalement en ce qui concerne les pronoms personnels. On remarque également que les corpus italiens contiennent moins de pronoms personnels à cause du phénomène de pro-drop [4]. Le nombre de pronoms personnels est plus élevé en anglais et en allemand parce qu'il n'existe pas de clitiques dans ces langues. Quant aux clitiques, l'italien présente des pourcentages plus élevés que le français parce qu'il y a plus de verbes pronominaux en italien et parce qu'on utilise des tournures verbales pronominales là où le français utilise le passif (Russo *et al.*, 2011). Au vu de ces données, on observe que les pronoms prennent une place considérable dans un corpus et que leur traduction constitue donc un problème à ne pas négliger.

3 Les pronoms dans la langue cible

Pour évaluer la qualité des traductions, nous avons fait traduire les deux corpus par Its-2 (Wehrli *et al.*, 2009), traducteur à base de règles linguistiques fondé sur l'analyseur syntaxique Fips. À titre de comparaison, nous avons également traduit les cinq cent premières phrases du corpus du *Petit Prince* [5] par Google Translate, traducteur à base statistique disponible en ligne [6]. Nous avons considéré les cinq paires de langues suivantes : allemand-français ; français-anglais ainsi qu'anglais-français ; français- italien ainsi qu'italien-français. Les traductions des pronoms ont été ensuite évaluées manuellement. En particulier, parmi les pronoms mal traduits, nous distinguons plusieurs cas de figure : i) mauvais type de pronom (1a) ; ii) forme pronominale incompatible avec la référence anaphorique (1b) ; iii) mauvaise position du pronom (1c) ; iv) ajout d'un pronom non requis dans la langue cible (1d) ; et v) absence d'un pronom requis dans la langue cible (1e).

1. Les textes en français, anglais et allemand sont disponibles sur `http://wikilivres.info` ; celui en italien sur `http://www.macchianera.net/files/ilpiccoloprincipe.pdf`.
2. Il s'agit des communiqués de 2007, tels que disponibles à l'addresse : `http://www.news.admin.ch`. Ce corpus est désormais appelé corpus de Presse ou Presse tout court.
3. Fips considère comme pronoms personnels à la fois les pronoms personnels forts (*moi* et *lui*) et les clitiques sujet (*je* ou *il*).
4. On définit le phénomène du pro-drop comme l'absence du pronom personnel sujet, qui peut ne pas être exprimé en italien grâce à la richesse de la morphologie verbale.
5. Ce choix a été dicté par le nombre considérable de pronoms présent dans le texte littéraire par rapport au texte de la presse.
6. `http://translate.google.com`

(1) a. Qui parle ?
* That speaks ?
b. L'homme et son chien.
* The man and her dog.
c. Il me répond.
* Risponde mi.
d. Il s'est suicidé.
* He committed himself suicide.
e. Je savais le reconnaître.
* Knew to recognize it.

Dans la Table 2, nous schématisons les résultats obtenus en termes de pourcentages de pronoms bien traduits, mal traduits et de pronoms manquants dans la langue cible.

	Its-2 Presse			**Its-2 Petit Prince**			**Google Petit Prince**		
	Bon	Mauvais	Manquant	Bon	Mauvais	Manquant	Bon	Mauvais	Manquant
EN-FR	54%	42%	4%	73%	24%	3%	93%	5%	2%
IT-FR	36%	**43%**	**20%**	51%	24%	26%	66%	18%	16%
DE-FR	42%	53%	5%	73%	21%	6%	75%	18%	7%
FR-EN	52%	43%	5%	76%	17%	7%	91%	5%	4%
FR-IT	57%	29%	14%	65%	26%	8%	58%	**21%**	**20%**

TABLE 2 – Comparaison des paires de langues. Les corpus évalués, traduits par Its-2, contiennent environ 1000 phrases (Presse) et environ 1600 phrases (*Petit Prince*). À titre d'illustration, on a évalué 500 phrases du *Petit Prince* traduites par Google Translate.

Globalement, ces résultats montrent que le pourcentage d'erreurs en traduction automatique est assez élevé, pouvant atteindre des valeurs de 63% (pronoms mal traduits et manquants dans Its-2 Presse, pour la paire italien-français) ou de 41% (dans Google Petit Prince, pour la paire français-italien). Le corpus de Presse est plus difficile à traduire, avec des baisses de performance entre 7% et 31% par rapport au corpus Petit Prince. Cette tendance est confirmée par une mini-étude annexe, non présentée ici, pour Google Translate, avec une baisse de performance du même ordre de grandeur pour le corpus Presse par rapport au corpus littéraire.

Cela confirme en partie l'affirmation de Kay (1986), en suggérant qu'il y a encore beaucoup à faire en traduction automatique des pronoms. Une difficulté majeure est causée par la différence entre les systèmes pronominaux de deux langues. Considérons, par exemple, la paire italien-français. À cause du pro-drop en italien, le système de traduction doit être capable de générer le pronom personnel sujet en français. Si ce n'est pas le cas, la phrase cible en français sera agrammaticale (2). Cela explique en partie les pourcentages élevés de pronoms manquants atteint par les deux systèmes.

(2) [...] e dormono durante i sei mesi che la digestione richiede.
* [...] et dort pendant les six mois que la digestion demande. (Its-2)
* [...] et dormir pendant les six mois qui nécessite la digestion. (Google Translate)
'*[...] et ils dorment pendant les six mois que la digestion demande.*'

Une autre difficulté réside dans la traduction des pronoms clitiques. Leur mauvaise traduction, voire l'absence de leur traduction, dans la langue cible est une erreur fréquente pour les traducteurs automatiques, comme déjà montré par les travaux de Russo (2011). Un exemple de ce problème est fourni en (3), où – dans la traduction du français vers l'italien – les deux systèmes ne traduisent pas le clitique *lui* (*gli* en italien).

(3) Alors je ne lui parlais ni de serpents boa, ni de forêts vierges, ni d'étoiles.
* Allora io parlavo né d' serpenti boa, né di foreste vergini, né di stelle. (Its-2)
* Allora non avrei mai parlare o boa constrictor, o primordiali foreste, o le stelle. (Google Translate)
'*Allora non gli parlavo nè di boa constrictor, nè di foreste vergini, nè di stelle.*'

4 Évaluation des résultats en Its-2

Pour mieux comprendre les difficultés que notre système rencontre lors de la traduction des pronoms, nous avons analysé dans les détails la traduction effectuée par Its-2 sur les deux corpus. Un avantage considérable d'un système de traduction automatique linguistique comme Its-2 est qu'il fournit directement les étiquetages morpho-syntaxiques des mots de la phrase cible. Un premier résultat constitue donc la distribution des catégories pronominales dans les différentes langues cibles (Table 3).

En particulier, les paires de langues ayant le français comme langue cible sont instructives : comme il s'agit des mêmes textes, la distribution des pronoms devrait être similaire. Or, beaucoup moins de pronoms apparaissent en italien-français (2982 occurrences, Table 3) que dans les deux autres paires de langues (3910 occurrences en anglais-français). Cela est dû à l'omission du pronom personnel sujet et des pronoms clitiques en français – comme déjà montré en (2) et en (3) –, mais aussi à la non-génération en français des pronoms démonstratifs *c'* et *ça* (4). On remarque aussi que, globalement, on génère davantage de pronoms démonstratifs en partant du français (0,99% en français-italien) qu'en traduisant vers le français (0,23% en italien-français). Ceci est à nouveau en rapport avec les démonstratifs faibles *c'* et *ça*, qu'on retrouve plutôt sous forme de pronom personnel neutre en allemand (*es*) et en anglais (*it*). En ce qui concerne la paire anglais-français, il peut arriver que l'analyseur syntaxique détecte des phrases relatives à pronom relatif nul de manière erronnée. Par conséquent, Its-2 génère des pronoms relatifs dans la langue cible alors qu'il ne s'agit pas d'une phrase relative.

(4) È il mio aeroplano.
* Mon avion est. (Its-2)
'*C'est mon avion.*'

Paire	CLI	PERS	POSS	DEM	REL	INDEF	INTER	Total	Pronoms	Mots
EN-FR	1,73%	4,08%	1,2%	0,45%	0,82%	0,34%	0,33%	8,96%	3910	43654
IT-FR	1,58%	3,76%	0,79%	**0,23%**	0,08%	0,09%	0,68%	7,21%	2982	41337
DE-FR	1,98%	4,39%	0,94%	0,53%	0,28%	0,39%	0,81%	9,31%	3450	37049
FR-EN	0%	4,47%	1,05%	0,94%	0,66%	0,16%	0,2%	7,47%	3135	41954
FR-IT	1,71%	3,8%	1,1%	**0,99%**	0,07%	0,15%	0,75%	8,57%	3394	39611

TABLE 3 – Distribution des pronoms générés par Its-2 dans la langue cible. Les chiffres représentent des pourcentages de tous les mots dans les corpus *Petit Prince* et Presse cumulés.

	CLI	PERS	POSS	DEM	REL	INDEF	INTER	Moyenne
EN-FR	77%	79%	81%	76%	75%	84%	61%	78%
IT-FR	75%	69%	68%	75%	75%	71%	60%	69%
DE-FR	52%	45%	52%	52%	50%	43%	29%	46%
FR-EN		71%	64%	74%	58%	75%	65%	69%
FR-IT	70%	70%	64%	74%	56%	66%	60%	69%

TABLE 4 – Pourcentage de pronoms évalués comme corrects pour chaque type de pronom.

Nous avons déjà montré les chiffres globaux de notre évaluation manuelle (Table 2). En distinguant les différentes catégories des pronoms, on obtient les chiffres détaillés de la Table 4. Globalement, on voit que la paire allemand-français est en retrait par rapport aux autres (à l'exception des pronoms relatifs en français-italien et en français-anglais). Cela est dû principalement à la plus grande distance linguistique entre ces deux langues. [7]

5 Pistes de recherche

L'objectif de notre recherche étant l'amélioration de notre système de traduction automatique sur les pronoms, nous avons repéré les deux principales difficultés qui subsistent pour notre système : l'ambiguïté des formes pronominales dans la langue source et l'ambiguïté du transfert vers la langue cible. Nous avons donc identifié deux pistes de recherche principales pour résoudre ces problèmes : la première consiste à améliorer la détection des pronoms, pour distinguer, par exemple, les formes homographes telles que les déterminants et les pronoms relatifs en allemand (*der, die, das, den, ...*) [8]. La deuxième piste concerne l'intégration d'un module d'identification des références pronominales, en suivant l'exemple de travaux tels que celui de Le Nagard & Koehn (2010) sur un système de traduction automatique statistique (TAS).

7. Pour une étude portant sur la mesure de la distance linguistique ainsi que sur leur influence sur la traduction automatique, voir Birch *et al.* (2008).

8. Dans ce cas spécifique, le pronom relatif en allemand est obligatoirement précédé d'une virgule ou d'une préposition elle-même précédée d'une virgule. Des contraintes supplémentaires dans l'analyse pourraient donc réduire cette ambiguïté.

Résolution des ambiguïtés d'analyse Une évaluation supplémentaire de l'étiquetage au niveau de la langue source montre que le pourcentage de pronoms mal étiquetés peut être assez élevé. Les chiffres se situent entre 2,74% (*Petit Prince* anglais), pouvant aller jusqu'à 34,27% (Presse allemand). [9] Parmi ces erreurs, on trouve des cas d'étiquetage de la mauvaise partie du discours. Au-delà de l'exemple allemand mentionné ci-dessus, on peut citer l'ambiguïté de formes telles que *la*, *le* et *li* en italien, qui peuvent être des déterminants ou des pronoms clitiques. On trouve également des cas d'étiquetage dans la mauvaise catégorie pronominale. Par exemple, *qui* en français peut être analysé comme un pronom interrogatif ou par un pronom relatif, tout comme *che* en italien (Russo *et al.*, 2011).

Des erreurs d'étiquetage plus fines, non comptabilisées dans les chiffres ci-dessus, concernent l'étiquetage morphologique. Par exemple, le pronom personnel allemand *sie* est soit singulier féminin, soit pluriel (tous genres). [10] En cas de pronom sujet, il peut être désambiguïsé facilement à l'aide de l'analyse syntaxique. Nos résultats pour la paire allemand-français confirment aussi les résultats obtenus par Hardmeier & Federico (2010), qui ont étudié la performance d'un traducteur automatique à base statistique sur les pronoms personnels pour la traduction de l'allemand vers l'anglais. Ils ont montré que les pronoms de politesse et les réflexifs, de par leur ambiguïté de genre et nombre, posent problème lors de la traduction.

Un autre exemple concerne l'ambiguïté du pronom *il* en français, qui peut être un pronom personnel ou impersonnel selon la construction syntaxique dans laquelle il apparaît. S'il apparaît comme pronom personnel, il faudra rechercher son antécedent pour pouvoir le traduire correctement dans la langue cible. S'il apparaît comme pronom impersonnel, la recherche de son antécedent dans le texte est inutile. [11] En ce qui concerne l'italien, on peut souligner encore l'ambiguïté de la forme *loro*, qui peut être un pronom personnel sujet de troisième personne du pluriel, mais aussi un pronom faible et un déterminant possessif. La distinction entre ces différentes formes peut être faite en fonction du groupe verbal ou du groupe nominal.

Résolution des anaphores pronominales Sur notre corpus, nous avons aussi spécifié quelles erreurs étaient dues à l'absence d'une résolution anaphorique. [12] En ce qui concerne le corpus de Presse, les erreurs d'anaphores occupent une partie importante. Par contre, dans le corpus du *Petit Prince*, les erreurs d'anaphores sont rares (Table 5), sauf en italien-français et en français-italien, où certains mots très fréquents dans le *Petit Prince* montrent une différence de genre : c'est le cas de *planète* et *fleur* (masculins en italien, féminins en français) ou *renard* (féminin en italien, masculin en français). En l'absence d'un module capable d'identifier les antécédents des pronoms, le système traduit un pronom féminin de la langue source par un pronom féminin de la langue cible. Ce problème est accentué par la surgénération du pronom personnel sujet en français-italien, non requis à cause du pro-drop. Les phrases générées restent grammaticalement correctes, mais dans certains cas la référence anaphorique n'est pas correcte.

	Mauvaise référence anaphorique	
	Petit Prince	Presse
EN-FR	1,58%	22,38%
DE-FR	1,34%	16,94%
IT-FR	6,27%	3,66%
FR-IT	8,19%	3,68%

TABLE 5 – Nombre de pronoms annotés comme faux à cause d'une référence anaphorique mal choisie.

Des systèmes de TAS à segments permettent une résolution d'anaphores basique : si la table de traduction contiennent des segments de longueur maximale n au niveau de la langue source, elle peut encoder des chaînes anaphoriques dont la distance entre le pronom et l'antécédent est inférieure ou égale à n. Les exemples en (5a-b),

9. Les erreurs d'étiquetage plus nombreux en allemand constituent une explication supplémentaire des mauvais résultats observés dans la Table 4.

10. Nous ne tenons pas compte ici d'une troisième possibilité, la forme de politesse *Sie*, qui ne peut être distingué syntaxiquement de la troisième personne du pluriel.

11. À ce propos, nous renvoyons le lecteur aux travaux de Danlos (2005) sur un système qui permet de récupérer les occurrences du pronom impersonnel *il*.

12. Nous n'avons pas fait cette annotation pour la paire français-anglais. Ce comptage concerne les erreurs, pas les cas de références anaphoriques en général. En effet, en utilisant une règle par défaut qui traduit l'anglais *it* par le français *il*, on arrive à résoudre un bon nombre de cas de manière correcte.

pris de nos expériences avec Google Translate, illustrent ce phénomène, où l'introduction d'un adverbe empêche le système de retrouver la chaîne anaphorique. À l'opposé, la distance de surface entre anaphore et antécédent n'est pas pertinente pour des systèmes de TA linguistiques.

(5) a. Elle choisissait avec soin ses couleurs. → her colors
b. Elle choisissait avec *énormément* de soin ses couleurs → its colors

Les chiffres présentés ci-dessus montrent que le problème de la résolution d'anaphores est fortement dépendant du registre du texte et que l'inclusion d'un tel module dans un système de TA – dans la ligne des travaux de Lappin & Leass (1994) et Mitkov *et al.* (2002) – pourrait sensiblement améliorer sa qualité. Même si des travaux récents sur l'inclusion d'un module de références anaphoriques dans un système de traduction automatique statistique n'ont pas montré de gains substantiels (Hardmeier & Federico, 2010; Le Nagard & Koehn, 2010), l'approche fondamentalement différente de notre système de TA linguistique pourrait plaider en faveur d'une telle extension.

6 Conclusion

Dans ce travail, nous avons montré l'importance des pronoms dans la traduction automatique. Après avoir établi la fréquence élevée des pronoms dans les corpus et dans les langues considérées, nous avons discuté les limites et les faiblesses des systèmes de traduction automatique et en particulier du nôtre, Its-2. Nous avons évalué dans les détails les traductions de notre système pour proposer quelques pistes de recherche qui devraient nous permettre d'améliorer ses performances.

Références

BIRCH A., OSBORNE M. & KOEHN P. (2008). Predicting success in machine translation. In *Actes de EMNLP 2008*, Honolulu.

DANLOS L. (2005). ILIMP : Outil pour repérer les occurrences du pronom impersonnel *il*. In *Actes de TALN'05*, p. 123–132, Dourdan.

HARDMEIER C. & FEDERICO M. (2010). Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, Paris.

KAY M. (1986). Machine translation will not work. In *Proceedings of ACL'86*, p. 286.

LAPPIN S. & LEASS H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**(4), 535–561.

LAURENT D. (2001). *De la résolution des anaphores*. Rapport interne, Synapse Développement.

LE NAGARD R. & KOEHN P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, p. 252–261, Uppsala.

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.

MITKOV R., EVANS R., ORĂSAN C., HA L. & PEKAR V. (2007). Anaphora resolution : To what extent does it help NLP applications ? In A. BRANCO, Ed., *Anaphora : Analysis, Algorithms and Applications* : Springer.

RUSSO L. (2011). La traduction automatique entre langues proches : les pronoms clitiques en italien et en français. In G. B. BARBEAU, C. GAGNÉ & G. LEBLANC, Eds., *Actes des XXIVes Journées de linguistique*, p. 141–153, Québec : CIRAL.

RUSSO L., SCHERRER Y., GOLDMAN J.-P., LOÁICIGA S., NERIMA L. & WEHRLI E. (2011). Étude interlangues de la distribution et des ambiguïtés syntaxiques des pronoms. In *Actes de TALN'11*, Montpellier.

TUTIN A. (2002). A corpus-based study of pronominal anaphoric expressions in French. In *Proceedings of DAARC 2002 (Discourse Anaphora and Anaphora Resolution)*, Lisbonne.

WEHRLI E. (2007). Fips, a "deep" linguistic multilingual parser. In *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*, p. 120–127, Prague.

WEHRLI E., NERIMA L. & SCHERRER Y. (2009). Deep linguistic multilingual translation and bilingual dictionaries. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, p. 90–94, Athènes.