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

Étude inter-langues de la distribution et des ambiguïtés syntaxiques des pronoms. *

Lorenza Russo, Yves Scherrer, 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

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

Résumé. Ce travail décrit la distribution des pronoms selon le style de texte (littéraire ou journalistique) et selon la langue (français, anglais, allemand et italien). Sur la base d'un étiquetage morpho-syntaxique effectué automatiquement puis vérifié manuellement, nous pouvons constater que la proportion des différents types de pronoms varie selon le type de texte et selon la langue. Nous discutons les catégories les plus ambiguës de manière détaillée. Comme nous avons utilisé l'analyseur syntaxique Fips pour l'étiquetage des pronoms, nous l'avons également évalué et obtenu une précision moyenne de plus de 95%.

Abstract. This paper compares the distribution of pronouns according to the text genre (literary or news) and to the language (French, English, German and Italian). On the basis of manually verified part-of-speech tags, we find that the proportion of different pronoun types depends on the text and on the language. We discuss the most ambiguous cases in detail. As we used the Fips parser for the tagging of pronouns, we have evaluated it and obtained an overall precision of over 95%.

Mots-clés : Pronoms, ambiguïté pronominale, étiquetage morpho-syntaxique.

Keywords: Pronouns, pronominal ambiguity, part-of-speech tagging.

1 Introduction

En traitement automatique du langage (TAL), la plupart des recherches sur les pronoms se sont focalisées sur la résolution des anaphores. Dans ce domaine, de très nombreux travaux traitent d'algorithmes capables de détecter des chaînes anaphoriques inter- et intra-phrastiques, de leur implémentation et de leur évaluation (Lappin & Leass, 1994; Mitkov *et al.*, 2002; Trouilleux, 2002). Beaucoup moins nombreux sont les travaux qui ont étudié l'impact de la résolution anaphorique sur des systèmes de TAL (Mitkov *et al.*, 2007; Hardmeier & Federico, 2010). Enfin, des études sur corpus ont été effectuées afin de quantifier la fréquence des pronoms anaphoriques dans différents types de texte (Tutin, 2002; Laurent, 2001).

Le travail que nous présentons dans cet article vise à répondre à trois questions plus générales, indépendantes du caractère anaphorique des pronoms : à quelle fréquence rencontre-t-on les pronoms dans les textes ? Est-ce que la distribution des pronoms change par rapport au type de texte et aussi par rapport à la langue utilisée ? Quelles sont les ambiguïtés pour chaque type de pronom ? Nous avons effectué une étude sur corpus afin d'étudier la distribution des pronoms dans deux textes différents (un texte littéraire et un corpus de communiqués de presse) et dans quatre langues (français, anglais, allemand et italien). Notre but principal est de mieux comprendre la distribution des pronoms en fonction du style du texte [1] et de la langue concernée et d'évaluer l'étiquetage de notre système

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

1. Nous préférons parler ici de style de texte plutôt que de genre de texte, car le texte littéraire que nous analysons ne peut pas être considéré comme un échantillon représentatif du genre littéraire.

d'analyse syntaxique. Un meilleur étiquetage syntaxique permettra également d'améliorer les performances de traducteurs automatiques à base de règles linguistiques (Scherrer *et al.*, 2011).

Cet article est structuré ainsi : dans les sections 2 et 3, notre étude plurilingue décrit la distribution des pronoms par langues, par style de texte et par catégorie grammaticale. À la section 4, nous détaillons les ambiguïtés pronominales inhérentes à chaque langue, alors que dans la section 5 nous discutons le problème des analyses incomplètes. La section 6 contient nos conclusions ainsi que de futures pistes de recherche.

2 Les corpus

Pour comparer la distribution des pronoms dans différents styles de texte et langues, nous avons choisi deux corpus disponibles dans quatre langues, à savoir un corpus littéraire – le *Petit Prince* en français, ainsi que ses traductions en anglais, allemand et italien [2] – et un corpus de communiqués de presse de l'Administration Fédérale Suisse [3], également disponible dans ces quatres langues.

Pour une telle étude, nous avions le choix entre des corpus annotés à la main – comme des banques d'arbres – et des corpus annotés automatiquement. Dans le premier cas, la qualité des annotations est garantie, mais le choix des documents est restreint. Il l'est encore plus lorsqu'on exige des corpus comparables dans plusieurs langues afin d'obtenir des généralisations à travers les langues. Dans le deuxième cas, la qualité d'annotation est moindre, mais le choix des textes est ouvert. Nous avons choisi d'étiqueter nos corpus automatiquement à l'aide de l'analyseur Fips (Wehrli, 2007) et de corriger manuellement ces annotations. Ces dernières nous permettent de discuter la distribution des pronoms en fonction du style du texte et de la langue, ainsi que d'évaluer la performance de Fips et de recenser les cas d'ambiguïtés les plus marquées.

Au total, nous avons vérifié et corrigé l'annotation automatique de 1000 phrases pour chaque langue, 500 du *Petit Prince* et 500 du corpus de Presse. En particulier, nous nous intéressons à toutes les têtes lexicales étiquetées par Fips comme pronoms personnels (*je, eux, nous*), [4] clitiques (*te, la*), démonstratifs (*ça, cela*), relatifs (*qui, que*), indéfinis (*chacun, personne*), interrogatifs (*qui*) et déterminants possessifs (*ma, leurs*).

3 Distribution des pronoms par corpus et par langue

La Table 1 montre la distribution des pronoms selon les langues et les corpus. Si on considère les nombres de pronoms présents dans les phrases annotées, les chiffres montrent qu'un texte peut contenir jusqu'à 17,6% de pronoms, soit presque un mot sur cinq. Leur fréquence change par rapport au style de texte, notre corpus littéraire présentant un nombre total de pronoms bien plus élevé que notre corpus de Presse. En effet, le *Petit Prince* contient de nombreux passages de dialogue et en conséquence beaucoup de pronoms de première et deuxième personne. A l'opposé, le nombre élevé de concepts différents introduits dans un communiqué de presse empêche l'utilisation massive de pronoms anaphoriques.

Le style du texte a aussi une influence sur la longueur des phrases et donc sur leur complexité. Si les phrases du *Petit Prince* contiennent en moyenne 10 mots, celles de la Presse contiennent environ 20 mots. La proportion de pronoms relatifs peut être prise comme un indice de la complexité d'une phrase (Table 2). Ainsi, dans le *Petit Prince*, jusqu'à 9% de tous les pronoms sont des pronoms relatifs (en italien). Dans le corpus de Presse, ce chiffre va jusqu'à 29,7% (en italien).

Pour ce qui est des pronoms personnels, par contre, leur proportion est plus élevée dans le *Petit Prince* que dans le corpus de Presse, atteignant un maximum de 69,1% pour l'anglais et de 72,1% pour l'allemand. En ce qui concerne l'italien, les pourcentages plus bas de pronoms personnels sont dus au phénomène du pro-drop, ou pronom sujet nul. En d'autres mots, en italien, le pronom personnel sujet peut ne pas être exprimé grâce à la richesse flexionnelle des verbes qui permettent de déterminer la personne et le nombre du sujet même quand celui-ci est absent.

2. Les textes en français, anglais et allemand sont disponibles sur `http://wikilivres.info`. Le texte en italien est disponible à la page : `http://www.macchianera.net/files/ilpiccoloprincipe.pdf`.

3. Il s'agit de phrases extraites des communiqués de presse de 2007, tels que disponibles à l'addresse : `http://www.news.admin.ch`. Nous appelons ce texte désormais corpus de Presse.

4. Fips considère comme pronoms personnels à la fois les pronoms personnels forts (*moi* et *lui*) et les clitiques sujet (*je* ou *il*).

Corpus	Phrases	Mots	Mots/Phrase	Pronoms	Pronoms/Mots
Petit Prince					
EN	500	5361	10,7	910	17,0%
FR	500	5109	10,2	891	17,4%
DE	500	5684	11,4	998	17,6%
IT	500	4553	9,1	479	10,5%
Presse					
EN	500	10399	20,8	259	2,5%
FR	500	11715	23,4	279	2,4%
DE	500	8968	17,9	192	2,1%
IT	500	11399	22,8	222	1,9%

TABLE 1 – Distribution des pronoms selon les langues et les corpus. Les chiffres des pronoms proviennent de l'annotation manuelle.

Corpus	CLI	PERS	POSS	DEM	REL	INDEF	INTER
Petit Prince							
EN		69,1%	15,6%	4,5%	4,1%	5,4%	1,3%
FR	19,5%	48,9%	11,1%	11,4%	5,3%	2,1%	1,6%
DE		72,1%	10,3%	5,6%	3,7%	6,7%	1,5%
IT	50,5%	13,4%	20,3%	4,0%	9,0%	0,8%	2,1%
Presse							
EN		39,0%	27,8%	7,3%	18,9%	6,9%	0,0%
FR	11,8%	24,7%	25,1%	11,5%	22,2%	3,9%	0,7%
DE		51,6%	21,4%	13,5%	9,9%	3,6%	0,0%
IT	35,6%	5,0%	20,3%	8,1%	29,7%	0,9%	0,5%

TABLE 2 – Pourcentages des différentes catégories de pronoms par rapport à l'ensemble des pronoms trouvés dans le texte.

Quant aux pronoms démonstratifs, ils atteignent un pourcentage assez élevé en français dans le corpus littéraire (11,4%) tout comme dans le corpus de Presse (11,5%). Cela est dû à l'utilisation de pronoms démonstratifs tels que *c'* et *ça* là où d'autres langues utiliseraient plutôt un pronom explétif (par exemple, *it* ou *there* en anglais). Mais les chiffres les plus saillants concernent les pourcentages très élevés de pronoms clitiques en italien (50,5% et 35,6%) par rapport au français (19,5% et 11,8%) dans les deux corpus. Ces chiffres sont en partie motivés par des raisons linguistiques et en partie par des raisons de style. En ce qui concerne les raisons linguistiques, il y a plus de verbes pronominaux en italien qu'en français (1a, b). Pour ce qui est des raisons de style, au lieu du passif classique en français, on utilise plutôt des tournures verbales pronominales en italien (1c, d).

(1) a. Non posso muover**mi**.
Je ne peux pas bouger.

b. **Ci** piace leggere.
Nous aimons lire.

c. 43 défauts ont été constatés.
Si sono constatati 43 difetti.

d. Il parlait d'événements survenus hier.
Parlava di avvenimenti verificati**si** ieri.

Les pourcentages obtenus pour le français peuvent être comparés avec ceux de Tutin (2002). Elle a notamment compté les expressions anaphoriques dans dix articles économiques du *Monde* (style de texte se rapprochant de notre corpus de Presse) et dans le roman *De la Terre à la Lune* de Jules Verne (style littéraire se rapprochant du *Petit Prince*). D'après ses comptages, entre 87% et 93,5% des expressions anaphoriques sont des pronoms personnels (clitiques ou pleins) et des déterminants possessifs. Même si ses définitions des classes de pronoms ne coïncident pas complètement avec les nôtres, nous obtenons des chiffres comparables : entre 61,6% et 79,5% des pronoms sont des pronoms personnels (clitiques ou pleins) et des déterminants possessifs (voir Table 2). Contrairement à nos corpus, Tutin ne trouve pas d'effet de style de texte dans le nombre de pronoms personnels et clitiques. Elle ne comptabilise pas les pronoms de première et deuxième personne, n'étant pas anaphoriques et n'étant pas aussi fréquents dans son corpus que dans d'autres textes littéraires.

4 Les ambiguïtés d'analyse

Dans les sections précédentes, la distribution des types de pronoms a été discutée sur la base de la révision manuelle de l'étiquetage automatique de Fips. Dans cette section, nous comparons les annotations automatiques de Fips aux révisions manuelles. La Table 3 montre la précision et le rappel pour les quatre langues. Elle doit être lue comme suit : la précision décrit le pourcentage de pronoms étiquetés correctement par Fips. Le rappel décrit le pourcentage de pronoms corrects trouvés par Fips. Par exemple, 86% des pronoms démonstratifs anglais détectés par Fips ont été confirmé par les annotateurs humains (précision). Parmi tous les pronoms démonstratifs trouvés par les annotateurs humains, 98% ont déjà été étiquetés comme tels par Fips (rappel).

Certaines catégories, comme les possessifs anglais ou les démonstratifs italiens (Table 3), ne posent aucun problème d'analyse, étant des formes non ambigües. En revanche, d'autres catégories posent plus de problèmes, comme les interrogatifs en italien avec seulement 16% de précision. Cela est principalement dû à des ambiguïtés linguistiques que Fips n'arrive pas à résoudre. Nous décrivons maintenant, langue par langue, ces cas problématiques.

	Anglais		**Allemand**		**Français**		**Italien**	
	Précision	Rappel	Précision	Rappel	Précision	Rappel	Précision	Rappel
CLI					96%	98%	96%	100%
PERS	97%	100%	99%	100%	93%	99%	99%	99%
POSS	100%	100%	97%	99%	99%	99%	96%	99%
DEM	86%	98%	76%	82%	100%	99%	100%	100%
REL	100%	90%	36%	89%	92%	84%	92%	84%
INDEF	86%	55%	72%	99%	92%	100%	67%	100%
INTER	55%	100%	83%	100%	33%	81%	16%	82%

TABLE 3 – Précision et rappel des étiquetages proposées par Fips, les deux corpus confondus.

En anglais, le mauvais rappel dans la catégorie des indéfinis (55%) est dû au pronom explétif *there* (par exemple, dans l'expression *there is...* 'il y a ...'). Fips l'étiquette comme un pronom personnel, alors que nous l'avons annoté plutôt comme un indéfini. Ici, il ne s'agit donc pas d'une ambiguïté à proprement parler, mais plutôt d'une divergence dans l'interprétation linguistique.

Le faible pourcentage de la précision des pronoms interrogatifs (55%) provient des pronoms indéfinis comme *nobody, anybody* 'personne', classés par Fips comme des pronoms interrogatifs. De même, *what* est étiqueté comme un interrogatif dans des phrases telles que (2) alors qu'il introduit une phrase relative.

(2) But he was in Turkish costume, and so *nobody* would believe *what* he said.
Mais il portait un costume turc, et *personne* n'avait cru *ce qu'*il disait.

Une dernière difficulté concerne le mot *that*, pour lequel Fips confond les fonctions de pronom relatif (90% de rappel) et de pronom démonstratif (86% de précision, 98% de rappel).

En allemand, il existe plusieurs cas d'ambiguïté entre déterminants et pronoms. Le déterminant défini (*der, die, das, den, ...*) (3a) est homographe au pronom relatif (3b), et les mêmes formes peuvent aussi être utilisées dans certains cas comme pronom démonstratif (3c). [5] Cela explique les pourcentages particulièrement bas des pronoms relatifs et démonstratifs (36% et 76% de précision, 89% et 82% de rappel).

(3) a. *Das* Bundesamt für Verkehr.
L' office fédéral des transports.
b. Ich habe ein Haus gesehen, *das* hunderttausend Franken wert ist.
J'ai vu une maison *qui* vaut cent mille francs.
c. *Das* ist ein Hut.
Ceci est un chapeau.

Quant au déterminant indéfini (*ein, eine, einen, ...*), certaines formes fléchies sont communes avec un pronom

5. Le pronom relatif en allemand est obligatoirement précédé d'une virgule ou d'une préposition elle-même précédée d'une virgule. Cet indice typographique pourrait être mis à profit dans une procédure de désambiguïsation.

indéfini. Fips a tendance à surgénérer des analyses comme pronom indéfini et à sous-générer des analyses comme déterminant. Cela explique des valeurs de précision basses (72%).

En français et en italien, la difficulté principale concerne l'ambiguïté orthographique des pronoms relatifs (4a, b) et des pronoms interrogatifs (4c) (*que* et *qui* en français, *che* en italien). [6] Les formes *que* et *che* peuvent encore être des conjonctions introduisant une phrase subordonnée (4d). Dans les deux langues, Fips surgénère des pronoms interrogatifs à la place de pronoms relatifs ou de conjonctions, d'où des chiffres de précision très bas (33% en français, 16% en italien).

(4) a. Il représentait un serpent boa *qui* digérait un éléphant.
Era il disegno di un boa *che* digeriva un elefante.
b. Voilà le meilleur portrait *que* j'ai réussi à faire de lui.
Ecco il migliore ritratto *che* riuscii a fare di lui.
c. *Que* fais-tu ?
Che fai ?
d. Crois-tu *qu'*il faille beaucoup d'herbe à ce mouton ?
Pensi *che* questa pecora necessiti di una gran quantità d'erba ?

5 Les analyses incomplètes

L'avantage principal de l'utilisation d'un analyseur syntaxique comme Fips pour l'étiquetage des pronoms est l'accès à des structures syntaxiques. En effet, si beaucoup d'étiqueteurs basent leurs décisions sur un contexte local (par exemple, des trigrammes de mots), Fips utilise la structure globale de la phrase pour l'étiquetage. Lorsque Fips n'arrive pas à construire un arbre syntaxique pour la phrase entière, il construit plusieurs arbres syntaxiques partiels et dérive les étiquettes morpho-syntaxiques à partir de ceux-ci. Évidemment, ces analyses incomplètes peuvent avoir un impact négatif sur la qualité de l'étiquetage. C'est sur cet aspect que nous nous focalisons dans cette section.

Plus précisement, nous considérons ici toutes les phrases qui ont été analysées complètement. Ceci correspond à approximativement 80% des phrases du corpus pour l'anglais, le français et l'italien, et à 60% des phrases pour l'allemand. Les chiffres que nous présentons dans la Table 4 se réfèrent donc à ce sous-ensemble du corpus.

	Anglais		**Allemand**		**Français**		**Italien**	
	Précision	Rappel	Précision	Rappel	Précision	Rappel	Précision	Rappel
CLI					99%	97%	98%	100%
PERS	98%	100%	99%	100%	95%	100%	100%	98%
POSS	100%	100%	100%	98%	100%	100%	99%	100%
DEM	95%	97%	92%	86%	100%	100%	100%	100%
REL	100%	86%	60%	100%	95%	100%	81%	93%
INDEF	96%	69%	90%	97%	100%	100%	67%	100%
INTER	60%	100%	100%	100%	86%	86%	50%	80%

TABLE 4 – Précision et rappel des étiquetages proposées par Fips, en tenant compte uniquement des phrases analysées complètement.

On constate que la précision des étiquetages augmente de manière importante dans certaines catégories. En particulier, Fips trouve moins de pronoms interrogatifs et de pronoms relatifs et la précision de leur étiquetage est plus élevée (de 33% à 86% pour les interrogatifs français, de 16% à 50% pour les interrogatifs italiens, si l'on compare les Tables 3 et 4.). Dans le cas d'une analyse incomplète, Fips n'arrive souvent pas à identifier l'antécédent d'un pronom relatif. En conséquence, il préfère une lecture interrogative (par exemple pour *qui/que* en français) ou une lecture démonstrative (par exemple pour *der/die/das* en allemand).

Ces résultats suggèrent aussi que dans les phrases non analysées, beaucoup de mots étiquetés comme pronoms sont

6. Contrairement au français, qui a deux formes distinctes du pronom relatif en fonction de sujet (*qui*) et d'objet direct (*que*), en italien il n'existe qu'une seule forme (*che*) pour les deux cas. Pour une étude plus détaillée des différences linguistiques entre ces deux langues, voir par exemple Arcaini (2000).

en réalité des conjonctions, comme déjà expliqué dans l'exemple (4d). Une piste de recherches futures concernera donc l'amélioration des analyses afin de limiter ce type d'erreurs lors de l'étiquetage.

6 Conclusion

Le travail que nous avons présenté s'est donné comme but de répondre à trois questions concernant la distribution des pronoms. Nous avons étiqueté deux corpus différents dans quatre langues et nous avons trouvé que le nombre de pronoms varie selon le texte et selon la langue. En particulier, notre corpus littéraire est caractérisé par un nombre élevé de pronoms personnels, tandis que le corpus de Presse contient des phrases plus longues et plus complexes, avec des pronoms relatifs plus fréquents. Nous avons également discuté des erreurs d'annotation faites par l'analyseur syntaxique Fips. Il en est ressorti que certaines catégories de pronoms sont plus ambiguës que d'autres et que la complétude de l'analyse est un facteur crucial pour une désambiguisation correcte. Nous envisageons donc d'améliorer la qualité des analyses syntaxiques.

Certains cas d'ambiguïtés inter-catégories ont été aussi discutés. Il existe, cependant, des cas où un pronom peut avoir différentes lectures dans la même catégorie pronominale (ambiguïtés intra-catégories). Par exemple, à l'intérieur de ce que nous avons nommé pronoms personnels, on trouve en français à la fois le *il* impersonnel et le *il* anaphorique. [7] On ne fait pas non plus de distinctions entre les différentes interprétations du *sie* allemand (3e personne singulier du féminin, 3e personne pluriel, et forme de politesse). Pourtant, ces différences sont cruciales pour la résolution des anaphores et pour la traduction automatique. Elles vont faire l'objet d'une étude future.

De plus, nous aimerions examiner plus en détail l'impact du style de texte. D'une part, l'analyse d'autres textes littéraires nous permettra d'obtenir des généralisations concernant les différences de genre. D'autre part, il s'agira d'examiner s'il existe un effet de traduction, comme en partie déjà mis en évidence par les exemples en (1). On pourrait par exemple s'attendre à des différences dans la distribution de pronoms selon que l'original du texte est en français ou en allemand.

Références

ARCAINI E. (2000). *Italiano e francese. Un'analisi comparativa*. Turin : Paravia/Scriptorium.

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.

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.

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*, p. 179 – 190 : Springer.

SCHERRER Y., RUSSO L., GOLDMAN J.-P., LOÁICIGA S., NERIMA L. & WEHRLI E. (2011). La traduction automatique des pronoms. Problèmes et perspectives. In *Actes de TALN'11*, Montpellier.

TROUILLEUX F. (2002). A rule based pronoun resolution system for French. In *Proceedings of Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*, Lisbonne.

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.

7. Nous renvoyons, à ce propos, aux trauvaux de Danlos (2005) sur un système qui récupère les occurrences du pronom impersonnel *il*.