Correction automatique par résolution d'anaphores pronominales

Maud Pironneau, Éric Brunelle, Simon Charest
Druide informatique, 1435 rue Saint-Alexandre, bureau 1040, Montréal (Québec)
developpement@druide.com

Résumé. Cet article décrit des travaux réalisés dans le cadre du développement du correcteur automatique d'un logiciel commercial d'aide à la rédaction du français. Nous voulons corriger des erreurs uniquement détectables lorsque l'antécédent de certains pronoms est connu. Nous décrivons un algorithme de résolution des anaphores pronominales intra- et interphrastiques s'appuyant peu sur la correspondance de la morphologie, puisque celle-ci est possiblement erronée, mais plutôt sur des informations robustes comme l'analyse syntaxique fine et des cooccurrences fiables. Nous donnons un aperçu de nos résultats sur un vaste corpus de textes réels et, tout en tentant de préciser les critères décisifs, nous montrons que certains types de corrections anaphoriques sont d'une précision respectable.

Abstract. This article relates work done in order to expand the performance of a commercial French grammar checker. We try to achieve the correction of errors only detectable when an anaphora pronoun is linked with its antecedent. The algorithm searches for the antecedent within the same sentence as the pronoun as well as in previous sentences. It relies only slightly on morphology agreement, since it is what we are trying to correct, and uses instead information from a robust syntactic parsing and reliable cooccurrences. We give examples of our results on a vast corpus, and try to identify the key criteria for successful detection. We show that some types of corrections are precise enough to be integrated in a large scale commercial software.

Mots-clés : Correcteur, Anaphores, Pronom, Saillance, Approche multistratégique, Cooccurrences.
Keywords: Grammar checker, Anaphora, Pronoun, Salience, Multi-Strategy Approach, Cooccurrences.

1 Introduction

1.1 Objet des travaux

L'objectif est la correction de certains types d'erreurs, et non pas la reconnaissance générale des entités.

(1a) O : {*Alice*} n'a rien mangé ce *matin*. Je **l**'ai LEVÉ et **l**'ai EMMENÉ à la *garderie*.

La seconde phrase de l'exemple ci-dessus comporte deux erreurs d'accord. On ne peut détecter ces erreurs que lorsque l'on relie les deux pronoms *l'*, avec lesquels s'effectue l'accord des deux participes passés, au nom propre *Alice* de la première phrase, c'est-à-dire en résolvant les deux anaphores pronominales. Sachant qu'*Alice* est un prénom féminin, on obtient la phrase corrigée suivante :

(1b) C : [Alice] n'a rien mangé ce matin. Je **l**'ai LEVÉE et **l**'ai EMMENÉE à la garderie.

Dans le présent article, nous commençons par définir deux types de corrections qui requièrent la résolution des anaphores pronominales. Puis nous décrivons les algorithmes appliqués pour la résolution de ces anaphores intra- et interphrastiques afin de tenter d'effectuer chaque type de correction. Nous évaluons ensuite les résultats sur chacun de ces types, et nous montrons que l'un d'eux est de qualité suffisante pour avoir fait l'objet d'une implantation commerciale.

1.2 Conventions d'écriture des exemples

Voici les conventions graphiques utilisées dans les exemples de cet article :

- en *italique* : les antécédents potentiels selon l'algorithme ;
- [entre crochets] : l'antécédent sélectionné par l'algorithme ;
- {entre accolades} : l'antécédent réel ;
- en **gras** : les pronoms ;
- en MAJUSCULES : le mot à corriger ou corrigé par la résolution des anaphores.

Les mentions « O : » et « C : » introduisent en alternance le texte original et le texte corrigé.

2 Description des corrections recherchées

Le correcteur effectue déjà des corrections liées à des pronoms à l'intérieur d'une phrase lorsque les contraintes syntaxiques le permettent. Il reconnait ainsi les antécédents des pronoms qui sont syntaxiquement liés à leur antécédent, tels que les pronoms relatifs (ex. 2) ou les pronoms réfléchis (ex. 3). Il corrige aussi le pronom lui-même dans certains cas précis, comme dans une reprise du sujet (ex. 4). D'autre part, sans en connaitre l'antécédent, un pronom complément d'objet direct (COD) à gauche du participe passé possédant un nombre intrinsèque provoque l'accord en nombre dudit participe (ex. 5).

(2) C'est la {[*lettre*]} **que** tu as ENVOYÉ à Alice. (ENVOYÉE)

(3) {[*Alice et Élise*]} **se** sont SERRÉS l'une contre l'autre. (SERRÉES)

(4) {[*Alice*]} a-t-**IL** été sage aujourd'hui ? (ELLE)

(5) Elle **les** a VU. (VUS)

La nouveauté recherchée est d'apporter des corrections de genre ou de nombre dépendant d'un antécédent non syntaxiquement lié à un pronom, repéré dans le texte comme étant le plus probable. Dans l'exemple 5, le correcteur considère *Elle les a vus* comme sans erreur, mais, s'il parvient à déceler l'antécédent du pronom *les* et que celui-ci est féminin, il pourra alors plutôt corriger pour *vues*.

2.1 Pronoms considérés

Nous nous sommes concentrés sur les pronoms dont les référents sont généralement exprimés dans le texte, soit les pronoms personnels sujets et COD de 3^{e} personne : *il/elle/ils/elles/le/la/les/l'*. Nous avons exclu les pronoms ayant souvent des référents déictiques, tels que les pronoms de 1re et de 2^{e} personne, les pronoms démonstratifs et les pronoms possessifs. Nous avons aussi exclu les pronoms ayant des référents par défaut, tels que les pronoms indéfinis (ex. : *certains*, *tous*, etc.). Les pronoms *en, y* et *on* ne nous intéressent pas ici puisqu'ils n'entrainent aucune correction (les accords ne sont pas obligatoires et sont discutables).

2.2 Choix des corrections les plus vraisemblables

Après des essais initiaux, nous avons décidé de restreindre les corrections par résolution d'anaphores aux cas où les erreurs sont les plus vraisemblables. Par exemple, le scripteur omet plus facilement d'appliquer des accords s'ils ne sont pas marqués phonétiquement, et, comme le note (Sauvageot, 1972), l'accord en nombre est souvent silencieux (ex. 8). Une correction inaudible est donc plus vraisemblable qu'une autre qui modifie la prononciation.

Outre l'audibilité, la vraisemblance d'une erreur dépend aussi de la graphie d'un nom et de sa fréquence d'utilisation. En effet, des mots commençant par une voyelle peuvent causer des confusions de genre, particulièrement s'ils sont peu utilisés. On peut vérifier la probabilité de ces confusions par une simple recherche comparative sur Internet (ex. 6 et 7).

(6) Un anagramme ou une anagramme ? (40 200 résultats pour la recherche "*un anagramme*" et 567 000 résultats pour la recherche "*une anagramme*" le 17 février 2014 ; c'est-à-dire environ 7 % d'erreurs.)

(7) Un octave ou une octave ? (19 700 résultats pour la recherche "*un octave*" et 430 000 résultats pour la recherche "*une octave*" le 17 février 2014 ; c'est-à-dire 4 % d'erreurs.)

Notons que cette erreur est sensible à la géographie. Par exemple, on trouve 2 400 résultats sur Google pour *une autobus* sur des sites canadiens (recherche "*une autobus*" *site:ca*), alors qu'on n'en trouve que 1 280 occurrences sur des sites français (dont beaucoup de bruit de passages québécois). Nous avons donc tenu compte de l'origine du locuteur, fournie par le logiciel, pour ces corrections.

En nous basant sur ces critères de vraisemblance, nous avons décidé de tenter de corriger les accords liés aux pronoms s'ils sont inaudibles et de corriger le pronom lui-même en nombre si le nombre est inaudible, et de le corriger en genre si l'antécédent est jugé d'un genre problématique. On obtient les deux grandes classes de corrections décrites dans les points 2.3 et 2.4.

2.3 Correction en genre des accords avec les pronoms COD LES/L'

(8) O : Ces {*fleurs*} viennent de mon *jardin*. Je **les** ai CUEILLIS pour vous.
C : Ces [fleurs] viennent de mon jardin. Je **les** ai CUEILLIES pour vous.

(9) O : {*Alice*} n'a rien mangé ce *matin*. Je **l**'ai LEVÉ et **l**'ai EMMENÉ à la garderie.
C : [Alice] n'a rien mangé ce matin. Je **l**'ai LEVÉE et **l**'ai EMMENÉE à la garderie.

2.4 Correction du genre et du nombre des pronoms IL/ILS/ELLE/ELLES et LE/LA

(10) O : Je suis allée voir mes deux {*grands-mères*}. Incroyable, comme **ELLE** PARLE fort !
C : Je suis allée voir mes deux [grands-mères]. Incroyable, comme **ELLES** PARLENT fort !

(11) O : Tu peux m'aider avec cet {*anagramme*} ? **IL** n'est vraiment pas facile.
C : Tu peux m'aider avec cette [anagramme] ? **ELLE** n'est vraiment pas facile.

(12) O : Cet {*anagramme*}, je ne **LE** vois pas très clairement.
C : Cette [anagramme], je ne **LA** vois pas très clairement.

3 Système de reconnaissance de l'antécédent

Notre système doit être robuste et de couverture large puisque le correcteur est appelé à analyser n'importe quel type de texte : journaux, récits, dialogues, courriels, rapports, etc. Ces contraintes influeront sur l'ensemble du processus de sélection de l'antécédent.

La question de choisir ou non un système de détection à base de classifieurs entrainés par apprentissage automatique s'est naturellement posée. Nous étions attirés par la simplicité, la clarté et la flexibilité d'un système à base de règles qui avait déjà fait ses preuves (Trouilleux, 2002). La campagne d'évaluation de système de coréférences CoNLL-2011 a confirmé notre décision. En effet, le système de (Lee et coll., 2011) à base de règles et de *tamis*, et non à base de classification et d'apprentissage automatique, a obtenu les meilleures performances.

3.1 Survol de l'algorithme

Nous avons ainsi opté pour la même approche que (Lappin et Leass, 1994), (Mitkov 1998) et (Trouilleux, 2002), soit un algorithme à base de règles utilisant une liste d'antécédents potentiels pondérés, avec deux différences notables. Premièrement, avant de lier à un pronom la liste de ses antécédents potentiels, nous tentons de trouver l'antécédent idéal

dans la phrase même où se trouve le pronom par un système de motifs syntaxiques. Deuxièmement, alors que les algorithmes de (Lappin et Leass, 1994) et (Mitkov 1998) procèdent en deux temps, c'est-à-dire extraction des antécédents puis évaluation des couples, nous y ajoutons une troisième étape : évaluation de la probabilité de la correction selon les couples présents. Cette dernière étape est du même type que celle apportée par le système CogNIAC (Baldwin, 1997), qui sélectionne les antécédents selon le nombre de couples proposés. Voici les grandes étapes de notre système :

Pour chaque phrase :
- analyse syntaxique complète, puis sélection et extraction des antécédents *a* potentiels (point 3.3) ;
- sélection des pronoms candidats à la résolution pronominale (point 3.4).

Pour chaque pronom *p* repéré comme candidat à la résolution :
- recherche d'un antécédent local par motifs syntaxiques à partir de *p* (point 3.5) ;
- si les motifs échouent, récupération des antécédents extraits dans la phrase courante et dans les phrases précédentes puis formation des couples (*p*, *a*) ;
- évaluation des couples (*p*, *a*) selon des contraintes (point 3.6) ;
- évaluation des couples (*p*, *a*) selon des préférences (point 3.6) ;
- évaluation de la vraisemblance de la correction selon les couples (*p*, *a*) présents (point 3.7).

3.2 Contexte d'implémentation

Notre algorithme s'appuie sur plusieurs ressources fournies par le logiciel correcteur.

- Le dictionnaire. Le logiciel comporte un dictionnaire de plus de 120 000 mots, doté de traits sémantiques, morphologiques et syntaxiques. Nous savons ainsi si l'antécédent est humain ou non, s'il s'agit d'une entité comptable, entre autres ;
- L'analyseur. L'analyse est basée sur une grammaire de dépendances. Cette analyse est suffisamment robuste pour être fiable et pour retrouver des constructions complexes. Les règles dans la résolution des anaphores sont basées sur le résultat de l'analyse et nous considérons que l'analyse est toujours exacte ;
- Les régimes verbaux détaillés. L'analyseur syntaxique possède de nombreuses informations sur la formation des syntagmes verbaux. Ces informations sont très utiles pour vérifier la compatibilité sémantique d'un antécédent potentiel et d'un verbe ayant le pronom à résoudre comme complément ;
- Les cooccurrences (Charest, Brunelle, Fontaine et Pelletier, 2007). L'analyseur a été utilisé pour constituer la liste des combinaisons lexicales les plus fréquentes pour un mot selon sa fonction syntaxique. Nous évaluons grâce à cette liste la crédibilité de chaque antécédent potentiel dans la position syntaxique du pronom. Cette méthode a aussi été employée pour la traduction automatique par (Wehrli et Nerima, 2013), qui en démontrent les résultats positifs.

3.3 Sélection et extraction des antécédents potentiels

On parcourt chaque phrase pour y repérer chaque syntagme nominal (SN) potentiellement antécédent d'un pronom *p*. Notre algorithme se voulant interphrastique, il faut retenir à la volée un accès efficace à tous les antécédents potentiels au cas où on en aurait besoin, même si une phrase ne contient elle-même aucun pronom anaphorique. On repère donc les antécédents potentiels avant même de traiter les pronoms.

L'analyse syntaxique permet d'éliminer les SN qui présentent le contexte, c'est-à-dire les dates, certains circonstanciels de lieu, et autres. Ensuite, nous tentons de déterminer le degré de saillance de chaque antécédent potentiel en le pondérant, positivement ou négativement. Les règles de pondération sont autonomes, s'appliquent individuellement et sont sans ordre déterminant. Les règles les plus influentes examinent les positions syntaxiques. Elles sont hiérarchisées selon la théorie du centrage définie par (Grosz et Sidner, 1986) et décrite plus précisément par (Cornish, 2000), qui donne une liste de paramètres à considérer. Ainsi, les règles pondèrent selon la structure de la phrase (fonctions) et du SN (enchâssements), mais également selon la position du SN dans le discours (titre, début ou fin de paragraphe). Aussi, nous évaluons le degré d'actualisation du SN (déterminant, relative, etc.). Enfin, d'autres règles reconnaissent des constructions emphatiques figées comme *c'est à Paul que je parle*, et pondèrent en conséquence.

Chaque antécédent potentiel obtient ainsi un poids de saillance qui sera utile aux évaluations ultérieures. Ce poids est conservé dans une liste associée à la phrase. Par économie, un antécédent doit atteindre un seuil de poids minimal, en deçà duquel il n'est pas retenu. Nous avons fait une rapide évaluation de l'impact de ces règles dans la section 4.5.

Avant de passer à la prochaine étape, nous fusionnons certains antécédents pour en créer un nouveau afin de prévoir les cas où l'antécédent est discontinu (ex. 13). Ces antécédents ont un poids très minime, mais ils permettront d'éviter

certaines corrections indues.

(13) {*Marie*} appelle {*Jean*} et **ils** partent ensemble à la campagne.

3.4 Sélection des pronoms

L'algorithme traite les pronoms visés lors d'un deuxième parcours de la phrase. S'il existe plusieurs pronoms dans la phrase, nous les traitons un à un de gauche à droite. Deux filtres annulent la recherche d'antécédent pour un pronom :

– Le pronom est impersonnel (pronom *il* ; la reconnaissance des pronoms impersonnels a constitué un travail connexe codé au sein même de l'analyseur d'Antidote, mais non décrit par le présent article) ;
– Le pronom reprend un élément phrastique ou non nominal (pronoms *le* ou *l'* ; le pronom *la* n'est pas envisagé dans ce cadre selon notre principe de vraisemblance de l'erreur). Ces informations nous sont fournies par l'analyseur et les régimes verbaux.

Dans les deux cas, le pronom sera masculin singulier. S'il y a ambigüité (ex. 14 et 15), alors la recherche d'antécédent est tout de même lancée. L'étape d'évaluation de la correction (point 3.7) prendra en compte cet état de fait et nous ne corrigerons qu'en cas de présence d'une cooccurrence.

(14) **Il** restera toujours un perdant.

(15) Je **l**'ai vu.

3.5 Repérage d'un antécédent local par motifs syntaxiques

Une série de motifs syntaxiques est appliquée afin de tenter de trouver l'antécédent d'un pronom à l'intérieur de la phrase où il se trouve. Un « motif » est un modèle d'analyse d'une structure syntaxique où le pronom et son antécédent sont clairement liés. Ces motifs se basent sur la fonction du pronom, sur les éléments syntaxiques présents et sur l'actualisation des mots. Bien qu'euristiques, ces motifs permettent des repérages très surs et rapides. Nous en avons défini une dizaine pour les sujets et une autre dizaine pour les pronoms COD. Les exemples 16 et 17 en présentent deux. Dans le premier cas, nous avons remarqué que lorsqu'un sujet est partagé par deux verbes, et que le second verbe a pour COD un pronom, l'antécédent le plus sûr est le COD du premier verbe lorsqu'il existe. L'exemple 17 expose le cas où l'on trouve un pronom sujet dans une proposition conjonctive avant le verbe principal. Lorsque le sujet et le pronom sont de même nombre et de même genre, les deux mots sont préférablement reliés anaphoriquement.

(16) Les maitres allumèrent les {[*chandeliers*]} et **les** laissèrent bruler durant des heures.

(17) Dans ces circonstances, même si **elle** s'en défend, la {[*droite*]}, collectivement, n'a pas réellement intérêt à clarifier ses intentions.

3.6 Évaluation des couples pronom-antécédent

Lorsque les motifs échouent, nous formons la liste des antécédents potentiels. La liste des antécédents de la phrase en cours d'analyse est unie à celles des deux phrases précédentes, si elles existent, comme le préconise (Mitkov, 1998). Au passage, nous fusionnons les antécédents qui auraient été répétés dans des phrases différentes en gardant préférablement les caractéristiques de l'antécédent le plus proche, avec une augmentation du poids de saillance, le cas échéant. Chaque antécédent est pondéré selon sa distance par rapport au pronom.

Une première série d'évaluations élimine les antécédents non pertinents syntaxiquement selon le principe B de la théorie du liage (Chomsky, 1981). Ce principe dicte qu'un pronom de 3^{e} personne ne peut être lié à un syntagme présent dans son domaine de liage. L'analyse syntaxique fine de la phrase nous permet de l'appliquer très rigoureusement. Contrairement aux systèmes connus (Hobbs 1978, Lappin et Leass 1994, Dagan et Itai 1990, Baldwin 1997, Mitkov 1998, Trouilleux 2002), seuls les critères syntaxiques sont éliminatoires lorsqu'une correction est envisagée : les critères morphologiques deviennent de simples préférences.

Une deuxième série d'évaluations réajuste le poids des couples selon quatre classes de règles. Les deux premières classes sont les plus importantes et ont été les plus ardues à implémenter. Dans les deux cas, il faut simuler une phrase dans laquelle l'antécédent remplace le pronom et la soumettre à l'analyseur, puis examiner les résultats obtenus, sans pénaliser l'analyse d'un trop grand cout en temps.

– Règles sémantiques. Elles vérifient la compatibilité de l'antécédent par rapport aux éventuelles restrictions sémantiques du verbe qui porte le pronom (ex. 18). Dans l'exemple, la sémantique du verbe (*agresser*) demande un COD humain (*spectateur*) : l'antécédent non humain (*fauteuil*) peut donc être éliminé.
– Règles statistiques. Comme (Wehrli, 2013), nous utilisons les cooccurrences pour évaluer la force d'une anaphore. On peut mesurer, pour chaque antécédent, la fréquence de l'antécédent dans la position syntaxique du pronom (ex. 19).

(18a) Alors que la {[*spectatrice*]} était assise confortablement dans son *fauteuil*, le *metteur* en scène **l**'a AGRESSÉE avec des *images* effrayantes.
(18b) *Le metteur en scène agresse le *fauteuil*.
(18c) Le metteur en scène agresse la *spectatrice*.

(19) O : Pas de *problème* pour {*maman*}, *Max* **l**'a APPELÉ. <cooccurrence *appeler maman*>
C : Pas de problème pour [maman], Max **l**'a APPELÉE.

Les autres règles ne demandent qu'une comparaison, très rapide, plus simple, mais elles sont moins précises. Leur impact sur le poids de chaque couple anaphorique est plus modéré.

– Règles morphologiques. Elles vérifient la compatibilité morphologique non seulement avec le pronom, mais aussi avec tous les éléments de la phrase qui s'accordent avec le pronom. Dans l'exemple 20, on voit qu'il y a un accord non audible avec le participe passé (*trouvée*) ainsi qu'avec le premier attribut du COD (*jolie*), mais qu'il existe un accord audible avec le second (*courte*). Le poids de la morphologie sera plus fort dans ce cas.
– Règles syntaxiques. Ici, nous recherchons les parallélismes et pondérons selon les degrés de similitude. Dans l'exemple 21, on voit qu'on utilise deux fois le même verbe, avec la même distribution des arguments (nous tenons compte de la présence du semi-auxiliaire *vouloir*).

(20) C : J'ai vu la {[*robe*]} que tu as achetée à *Alice* ! Je **l**'ai TROUVÉE JOLIE, mais trop COURTE à mon gout...

(21) C : {[*Alice*]} ne voulait pas manger son {[*potage*]}, mais je l'ai forcée à y gouter. Bilan : **elle l**'a MANGÉ en deux minutes.

Pour déterminer les poids des règles, nous avons utilisé le corpus annoté ANANAS (Tutin et coll., 2000). Nous avons créé un outil de non-régression, c'est-à-dire un comparateur d'analyse pour chaque phrase du corpus, afin de repérer et de comptabiliser automatiquement les changements produits par la variation de poids d'une règle.

3.7 Évaluation de la correction et prise de décision

Dans le cas où l'antécédent ayant le poids le plus fort ne déclenche pas de correction, il est retenu : on donne le bénéfice du doute au scripteur. Bien que notre but ne soit pas de reconnaitre les antécédents des pronoms, ce résultat est tout de même présenté à l'utilisateur dans le correcteur sous la forme d'une note dans l'infobulle décrivant la nature et la fonction du pronom (figure 1). L'antécédent est aussi présenté dans l'analyse détaillée de la phrase (figure 2). Nous avons repris l'exemple 21 pour l'illustrer.

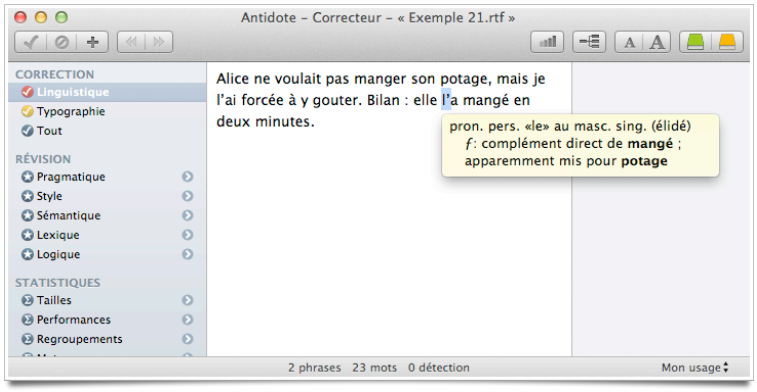

FIGURE 1 : Infobulle décrivant la nature et la fonction du pronom

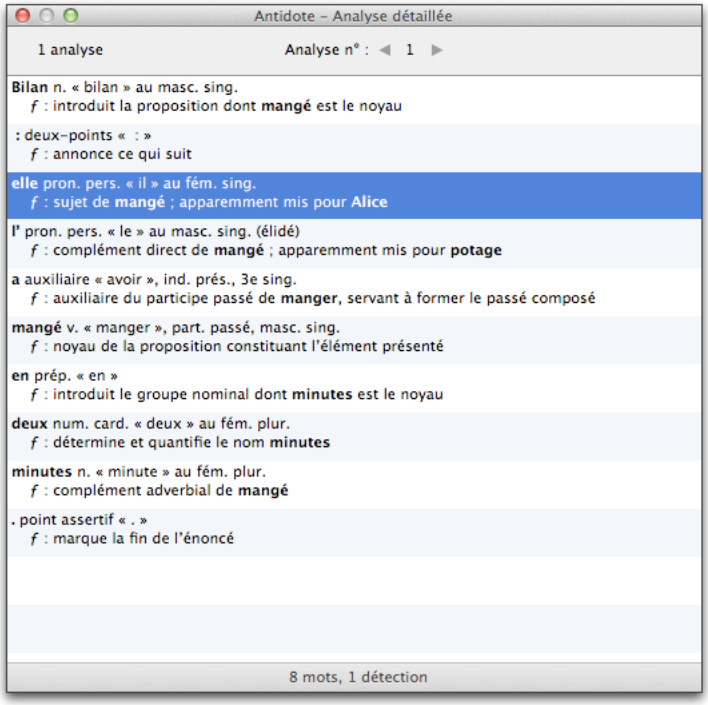

FIGURE 1 : Analyse détaillée de la phrase donnée par le logiciel de correction

Dans le cas où l'antécédent ayant le poids le plus fort déclenche une correction, nous entamons la procédure certainement la plus décisive du processus : l'évaluation de la vraisemblance de cette correction. Ce traitement prépondérant est largement euristique. Afin de filtrer le maximum de corrections indues, nous réévaluons globalement toute la situation du pronom et de son antécédent dans le texte. La sélection ne s'effectue plus selon le critère de poids du couple pronom-antécédent seul, même s'il reste important, mais selon des critères tels que la présence d'une cooccurrence entre le verbe dont le pronom dépend et l'antécédent, la différence de poids avec les autres antécédents disponibles ou la présence d'un antécédent moins fort mais sans correction. Nous vérifions aussi dans ce cadre si notre antécédent élu pourrait être lié plutôt à un autre pronom dans la phrase, ce qui éviterait la correction. Nous réévaluons aussi la correction selon l'éventualité que le pronom reprenne un élément phrastique ou soit impersonnel.

Dans le cas où nous hésitons sur une correction à apporter, car nous observons un autre antécédent possible n'apportant pas de correction ou parce que nous n'avons pas trouvé l'antécédent tout simplement, nous avons décidé d'« alerter » l'utilisateur en soulignant le pronom et en lui expliquant notre hésitation (figure 3). Le concept d'alerte est déjà utilisé pour mettre en avant des situations où l'attention de l'utilisateur est requise. Ces alertes sont soumises à un réglage qui permet à l'utilisateur de les inhiber s'il les trouve inopportunes.

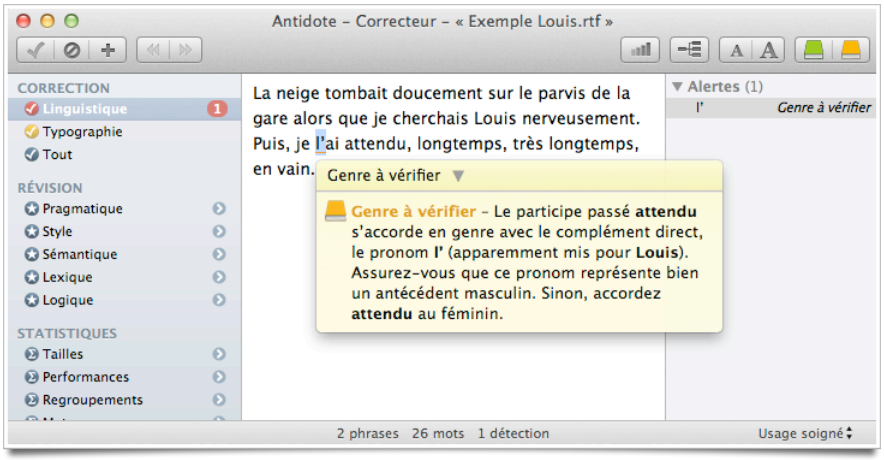

FIGURE 3 : Infobulle d'alerte

4 Résultats

4.1 Exemples de corrections obtenues

Voici quelques corrections obtenues lors de nos tests. Ces phrases sont de véritables phrases glanées sur Internet, retranscrites telles quelles. Les exemples 22 et 23 illustrent les corrections en genre (ex. 22) et en nombre (ex. 23) du pronom lui-même.

(22) O : Les {*incendies*} en Californie n'ont épargné *personne* se trouvant sur leur *chemin*. **ELLES** se sont ATTAQUÉ aux *monts* élevés comme au *fond* des *canyons*.
C : Les [incendies] en Californie n'ont épargné personne se trouvant sur leur chemin. **ILS** se sont ATTAQUÉS aux monts élevés comme au fond des canyons.

(23) O : La {*plupart*} des enfants éprouvent des difficultés à rester assis au moment des *repas* : **IL** se DANDINE, s'ASSOIT juste sur une *fesse*, se TORTILLE, se METTE à genoux…
C : La [plupart][1] des enfants éprouvent des difficultés à rester assis au moment des repas : **ILS** se DANDINENT, s'ASSOIENT juste sur une fesse, se TORTILLENT, se METTENT à genoux…

Les exemples 24, 25, 26 et 27 illustrent les corrections (toutes justifiées) effectuées sur des participes passés liés au pronom par instanciation du genre des pronoms COD *l'* et *les*. La figure 4 illustre le résultat dans le correcteur.

(24) O : J'ai l'{*épisode*}, mais je ne **l**'ai pas encore REGARDÉE.
C : J'ai l'[épisode], mais je ne **l**'ai pas encore REGARDÉ.

(25) O : D'après ces *sondages*, 18 % des *usagers* d'*appareils* numériques, ou 12 % des *foyers* américains, ont acheté une {*imprimante*} photo en 2006. Seuls *14* % d'entre eux **l**'ont ACHETÉ en kit avec un *appareil*.
C : D'après ces sondages, 18 % des usagers d'appareils numériques, ou 12 % des foyers américains, ont acheté une [imprimante] photo en 2006. Seuls 14 % d'entre eux **l**'ont ACHETÉE en kit avec un appareil.

[1] *La plupart*, bien qu'étant singulier, joue un rôle d'antécédent pluriel et doit être donc être repris par un pronom pluriel (Grevisse & Goosse).

(26) O : Soit, nous nous engageons dans une *construction* européenne rénovée qui respecte les {*Etats-Nations*} telles que l'Histoire **les** a FORGÉES.
C : Soit, nous nous engageons dans une construction européenne rénovée qui respecte les [États-Nations] tel que l'Histoire **les** a FORGÉS.

(27) O : Toutes ces {*actions*} de petite délinquance, que les *socialistes* appelaient les *incivilités*, Tony Blair **les** a APPELÉS comportements antisociaux.
C : Toutes ces [actions] de petite délinquance, que les socialistes appelaient les incivilités, Tony Blair **les** a APPELÉES comportements antisociaux.

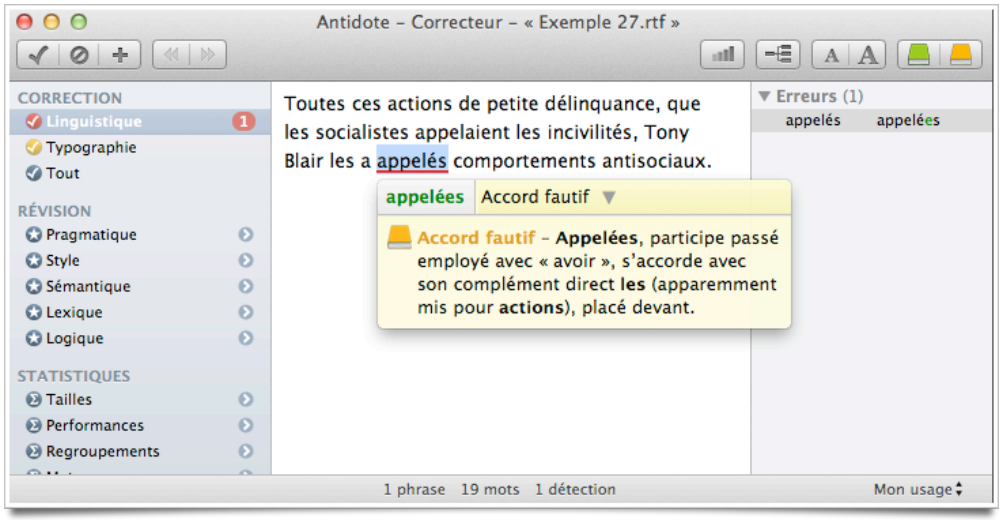

FIGURE 4 : Infobulle de correction de l'exemple 27.

4.2 Les contraintes de notre évaluation

L'évaluation classique de la résolution d'anaphores consiste à mesurer les taux de précision et de rappel de la reconnaissance de l'antécédent. Nous nous distancions de cette approche pour de multiples raisons. Tout d'abord, nous nous intéressons plutôt aux taux de précision et de rappel des corrections effectuées grâce à la résolution d'anaphores. D'autre part, l'évaluation se fait habituellement sur des textes considérés comme sans erreurs (journaux, etc.) ; dans notre cas, elle doit au contraire s'effectuer sur des textes bruts, avec de multiples erreurs. Enfin, nous ne faisons pas abstraction dans nos résultats des cas où l'erreur n'est pas due à notre système, mais plutôt à l'analyseur, qui n'aurait pas su, par exemple, faire la différence entre un pronom anaphorique et un mot d'une autre catégorie.

4.3 Évaluation de la précision

Évaluer le rappel de la correction est difficile, puisqu'il faudrait relever manuellement toutes les erreurs de référence de pronoms dans un corpus réel. Or, la configuration requise pour ce type de correction n'est pas si fréquente, et en trouver un nombre significatif serait un travail colossal. Nous avons ainsi fait le choix de n'évaluer que la précision de correction de l'algorithme, c'est-à-dire le nombre de bonnes corrections sur le nombre total de corrections.

L'évaluation de la précision nous a demandé elle aussi une réflexion, car trouver des cas de correction réels reste un défi en soi. Nous avons commencé en itérant sur un récit (un journal de voyage) augmenté d'un texte juridique, contenant plus de 20 000 phrases, afin de raffiner notre algorithme. Nous n'y avons trouvé aucun cas d'erreur qui nous intéressait. D'autre part, il nous fallait tester notre algorithme sur des textes de toutes origines, étant donné la diversité des utilisateurs du logiciel. Nous avons ainsi décidé d'utiliser le Web pour constituer un corpus plus vaste. Nous avons fait une sélection de sites selon plusieurs critères : révisés/non révisés ; caractère général/spécialisé ; présence de la 3e personne dans les pronoms personnels, objectif du site : informer/questionner/décrire/expliquer. Cette sélection comporte des textes descriptifs, informatifs, explicatifs, des biographies, des comptes-rendus de lectures, ainsi que des récits. De cette manière, nous avons extrait plus de 8 600 cas corrigés par notre algorithme. Ces 8 600 cas sont la somme des cas collectés en plus de 10 itérations sur des groupes de sites différents pour chaque itération. Notons qu'il

ne s'agit pas de 8 600 cas qui nécessitaient réellement des corrections : les cas récoltés lors de la première itération étaient très nombreux, mais étaient majoritairement des corrections injustifiées. Notre algorithme améliorant sa précision au fur et à mesure des itérations, le nombre total des corrections extraites a largement diminué, laissant une place plus importante aux corrections exactes.

Nous avons extrait finalement 152 cas lors d'une dernière itération sur de nouveaux sites ; ce sont ces cas qui nous ont servi de corpus d'évaluation, nous fournissant les chiffres du tableau 1. Sur les 152 cas, 107 corrections étaient exactes.

Types de correction	Nombre total de corrections	Nombre de corrections exactes	Précision
Correction du pronom lui-même[1]	29	18	62,0 %
Correction des éléments à accorder avec le pronom	123	95	77,2 %

TABLEAU 1 : Évaluation de la précision de la correction

L'évaluation a déterminé les types de correction à commercialiser. La partie de notre système qui corrige les éléments à accorder avec le pronom a été jugée de précision suffisante pour être commercialisée, tandis que l'autre partie devra continuer d'être améliorée.

4.4 Évaluation de l'espace mémoire et du temps de notre processus

Il est impératif de ne pas pénaliser l'utilisateur par un système de résolution trop gourmand en espace ou en temps. Après plusieurs ajustements techniques, nos résultats montrent finalement que la résolution des anaphores coute au correcteur 8 % de plus en temps et 5 % de plus en mémoire vive.

4.5 Évaluation de l'efficacité respective des critères utilisés

Nous avons fait quelques tests sur les 107 corrections exactes de notre corpus d'évaluation afin de montrer quels étaient les critères le plus importants dans la tentative de correction. Voici les trois faits les plus intéressants.

1. Sur notre corpus, 27 % des bonnes corrections sont effectuées par le biais des motifs. Donc, dans 1 cas sur 4, nous sommes en présence d'une construction relativement figée, où l'antécédent est présent dans la même phrase que le pronom. Un cas de correction indue relève des motifs (ex. 28). Il s'agit ici d'une erreur d'analyse où un syntagme nominal juxtaposé est analysé comme un élément mis en évidence (voir l'exemple 29 pour une phrase ayant la bonne analyse).

(28) O : Pourtant ce soir {*Fernando Manuel*} est humilié. Une vulgaire *histoire* de *femme*, une *portugaise* **l**'a VENDU.
C : Pourtant ce soir Fernando Manuel est humilié. Une vulgaire [histoire] de femme, une portugaise **l**'a VENDUE.

(29) La voiture, je l'ai vendue

[1] Seuls des pronoms personnels sujets ont été corrigés en nombre dans notre corpus (aucun pronom COD n'a été corrigé). Nous ignorons si cet état de fait est dû à la rareté de la correction ou à une mauvaise estimation de notre algorithme. Nous avons gardé cette question pour une évaluation future. Notons que la correction fonctionne pour des phrases avec des erreurs créées de toutes pièces.

2. En inhibant nos règles de saillance, nous perdons 31 % de nos corrections. De plus, plusieurs corrections inexactes apparaissent dans notre autre corpus de 20 000 phrases (ex. 30). La saillance a donc un impact majeur sur la correction par résolution d'anaphores.

(30) O : La {*Boule*} ne quitta pas des *yeux* l'alimentation. **IL** ÉTAIT INQUIET et ne COMPRENAIT pas pourquoi le niveau d'énergie ne montait pas.
C : La Boule ne quitta pas des [yeux] l'alimentation. **ILS** ÉTAIENT INQUIETS et ne COMPRENAIENT pas pourquoi le niveau d'énergie ne montait pas.

3. Enfin, en inhibant l'utilisation des cooccurrences, nous constatons qu'elles n'ont pas d'effet significatif sur la correction. Mais elles se montrent d'une grande efficacité dans la reconnaissance des pronoms clitiques COD pouvant reprendre un antécédent phrastique. Lorsque nous hésitons entre un antécédent nominal et un antécédent phrastique, en l'absence d'une cooccurrence, l'antécédent phrastique se révèle beaucoup plus probable. Notre résultat ne contredit pas (Wehrli et Nerima, 2013), mais cible l'aspect positif des cooccurrences dans le cadre de la résolution d'anaphores pour la correction automatique.

5 Conclusion

Cet article expose les caractéristiques de notre algorithme, lequel a la spécificité de corriger des erreurs grâce à la reconnaissance des liens anaphoriques sur un sous-ensemble de pronoms. Nous avons appliqué le principe de vraisemblance de l'erreur, et nous avons montré par nos résultats qu'il était bel et bien possible à l'heure actuelle de corriger précisément grâce à la référence pronominale intraphrastique et interphrastique. Nous avons vu qu'il est possible de mettre en place des solutions mélangeant plusieurs stratégies, telles que les motifs syntaxiques et la traditionnelle liste d'antécédents pondérés selon la méthode de (Lappin et Leass, 94), actualisée et précisée avec l'ajout des règles liées aux cooccurrences, sans que cela soit couteux en temps ou en mémoire.

Le code est maintenant testé et éprouvé quotidiennement par un demi-million d'utilisateurs. Quelques rares utilisateurs nous ont signalé une correction indue ou, au contraire, un silence, directement reliés à notre système. Notons qu'il est difficile de juger de la qualité du système sur ces requêtes, puisqu'il est rarissime qu'un utilisateur nous écrive pour nous féliciter d'une correction exacte. Pour notre part, en tout cas, nos propres textes ont bénéficié plusieurs fois de corrections reliées aux anaphores.

Nous avons depuis continué notre travail, et nous avons mis en place de façon connexe un système de motifs apportant des corrections nouvelles pour les pronoms de 2e personne ainsi que pour la correction des cas de pronoms reprenant des éléments phrastiques. Nous continuons aussi d'améliorer l'analyse syntaxique, renforçant de ce fait la qualité des corrections par les anaphores. Nous projetons de continuer notre travail sur la correction des pronoms eux-mêmes par l'ajout de nouvelles méthodes innovantes.

Remerciements

Merci à Mala F. Bergevin pour ses conseils linguistiques toujours avisés, et à Guy Lapalme pour ses suggestions concernant l'écriture de cet article.

Références

BALDWIN B. (1997). CogNIAC: A High Precision Pronoun Resolution Engine. Technical report, University of Pennsylvania.

CHAREST, S., BRUNELLE, E., FONTAINE, J., PELLETIER, B. (2007). Élaboration automatique d'un dictionnaire de cooccurrences grand public. Actes de *TALN 2007*, 283-292.

CHOMSKY, N. (1981). *Lectures on Government and Binding*, Dordrecht, Foris.

CORNISH, F. (2000). L'accessibilité cognitive des référents, le centrage d'attention et la structuration du discours : une vue d'ensemble. *Verbum*, Vol. XXII, no 1, 7-30.

DAGAN, I., ITAI, A. (1990). Automatic Processig of Large Corpora for Resolution of Anaphora References. Acte de 13th Conference on Computational Linguistics (COLING'90), 3, 330-332

GREVISSE, M., GOOSSE, A. (2007). *Le Bon Usage : grammaire française*, 14e éd., Bruxelles : De Boeck Duculot, 2008.

GROSZ, B., SIDNER, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 175-204.

HOBBS, J. (1978). Resolving Pronoun Reference. *Lingua* 44, 311-338.

LEE, H., PEIRSMAN, Y., CHANG, A., CHAMBERS, N., SURDEANU, M., JUFRAFSKY, D. (2011). Stanford's Multi-Pass Sieve Corefernce Resolution System at the CoNLL-2011 Shared Task. Acte de *CoNLL-2011 : Shared Task,* June, 73.

LAPPIN, S., LEASS, H. (1994). An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics* 20(4), 535-561.

MITKOV, R. (1998). Robust pronoun resolution with limited knowledge. Acte de *Annual Meeting of the ACL*, 869-875.

SAUVAGEOT, A. (1972). *Analyse du français parlé*. Paris : Hachette.

TROUILLEUX, F. (2002). A Rule-based Pronoun Resolution System for French. Acte de *4th Discourse Anaphora and Anaphor Resolution Colloquium*.

TUTIN A., TROUILLEUX F., CLOUZOT C., GAUSSIER E., ZAENEN A., RAYOT S., ANTONIADIS G. (2000). Annotating a large corpus with anaphoric links. Actes de *Discourse Anaphora and Anaphor Resolution (DAARC 2000)*.

WEHRLI, E., NERIMA, L. (2013). Collocations and anaphora resolution in machine translation. Acte de *Workshop on Multi-Word Units in Machine Translation and Translation Technologie*.