Confondre le coupable : corrections d'un lexique suggérées par une grammaire

Lionel NICOLAS[1], Jacques FARRÉ[1], Éric VILLEMONTE DE LA CLERGERIE[2]
[1] Laboratoire I3S, Université de Nice-Sophia Antipolis, CNRS
2000 route des Lucioles, B.P. 121, 06903 Sophia Antipolis Cedex, France
[2] Projet ATOLL - INRIA
Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France
{lnicolas,jf}@i3s.unice.fr,
Eric.De_La_Clergerie@inria.fr

Résumé. Le succès de l'analyse syntaxique d'une phrase dépend de la qualité de la grammaire sous-jacente mais aussi de celle du lexique utilisé. Une première étape dans l'amélioration des lexiques consiste à identifier les entrées lexicales potentiellement erronées, par exemple en utilisant des techniques de fouilles d'erreurs sur corpus (Sagot & Villemonte de La Clergerie, 2006). Nous explorons ici l'étape suivante : la suggestion de corrections pour les entrées identifiées. Cet objectif est atteint au travers de réanalyses des phrases rejetées à l'étape précédente, après modification des informations portées par les entrées suspectées. Un calcul statistique sur les nouveaux résultats permet ensuite de mettre en valeur les corrections les plus pertinentes.

Abstract. Successful parsing depends on the quality of the underlying grammar but also on the quality of the lexicon. A first step towards the improvement of lexica consists in identifying potentially erroneous lexical entries, for instance by using error mining techniques on corpora (Sagot & Villemonte de La Clergerie, 2006). we explores the next step, namely the suggestion of corrections for those entries. This is achieved by parsing the sentences rejected at the previous step anew, after modifying the information carried by the suspected entries. Afterwards, a statistical computation on the parsing results exhibits the most relevant corrections.

Mots-clés : analyse syntaxique, lexique, apprentissage, correction .

Keywords: parsing, lexicon, machine learning, correction .

1 Introduction

L'analyse syntaxique d'une langue repose sur l'utilisation de ressources linguistiques les plus précises et correctes possibles. Obtenir des ressources possédant une si large couverture est une tâche ardue de longue haleine qu'il est souhaitable d'alléger par le biais de techniques qui en automatisent l'élaboration et la correction. Nous présentons ici une technique de génération automatique de suggestions de corrections pour les entrées potentiellement erronées d'un lexique.

Nous nous intéressons aux moyens de réduire l'inexactitude et l'incomplétude d'un lexique à partir d'un recensement de formes lexicales suspectées d'être incorrectement ou seulement

partiellement décrites dans un lexique. Nous nous situons ainsi dans le prolongement direct de la technique de fouille d'erreurs sur des corpus de grande taille originalement proposée par (van Noord, 2004), et améliorée par (Sagot & Villemonte de La Clergerie, 2006). La pertinence de cette dernière s'observe notamment à travers nos résultats.

Cette technique repose sur l'idée suivante : étant donné un large corpus de phrases attestées, plus une forme (et indirectement les lemmes associés) apparaît ou n'apparaît pas dans des phrases dont les analyses échouent, plus nous avons des raisons de douter ou de ne pas douter des entrées lexicales qui lui sont associées. Cependant le contexte des formes importe : une forme est d'autant plus suspecte qu'elle apparaît dans des phrases non analysables mais en co-occurrence avec des formes qui tendent à apparaître dans des phrases analysables.

L'implémentation de la technique de fouilles d'erreurs nous a fourni une liste de 5344 formes suspectes avec, pour chaque forme f, un taux de suspicion et une liste de phrases non analysables (56089 au total) où f est suspectée d'être à l'origine de l'échec des analyses. Si une forme est effectivement responsable de ces échecs, et non la grammaire[1], c'est donc que les informations lexicales qui lui sont associées sont incomplètes ou inexactes (voir inexistantes).

En relâchant les contraintes sur les informations portées par une forme suspecte ou en les modifiant (notamment la catégorie syntaxique), de nouvelles analyses des phrases associées vont aboutir. Les représentations des phrases alors produites représentent les conditions dans lesquelles l'analyse a réussi, c.a.d. les informations sur la forme suspecte rendant possible l'analyse. En examinant ces informations sur un ensemble de phrases, il est alors possible de dégager des hypothèses de correction utiles.

La technique présentée est indépendante du langage étudié.

Travaux relatifs. L'acquisition de connaissances linguistiques depuis des corpus bruts (i.e. non annotés) par le biais de connaissances grammaticales a été initialement étudiée par (Brent, 1993) afin d'identifier les cadres syntaxiques des verbes en anglais. (Horiguchi *et al.*, 1995) utilisent les résultats d'analyse fournis par un système HPSG afin d'acquérir des entrées lexicales de mots japonais inconnus. Enfin, mentionnons la reconstitution d'informations lexicales manquantes en vue d'analyses robustes (Grover & Lascarides, 2001), (Crysmann *et al.*, 2002).

Nous commençons par expliquer comment générer des hypothèses de correction (Sect. 2) et comment les trier (Sect. 3). Nous introduisons ensuite la notion de synchronisation entre un lexique et une grammaire (Sect. 4), juste avant d'exposer les résultats obtenus (Sect. 5) et les développements futurs (Sect. 6).

2 Génération d'hypothèses

Le principal but d'un analyseur syntaxique est de vérifier la validité syntaxique d'une phrase et d'en produire une ou plusieurs représentations. On souhaite en général éviter la surgénération des représentations issues d'une analyse en produisant le moins possible de représentations.

Une phrase est qualifiée d'*ambiguë* pour un analyseur lorsque celui-ci lui associe plusieurs interprétations. Ceci arrive principalement lorsque la phrase est intrinsèquement ambiguë, i.e. d'autres informations (tel que le contexte sémantique) sont nécessaires afin de filtrer les inter-

[1]Nous supposons que les erreurs dues à un traitement incorrect en amont du processus d'analyse proprement dit (segmentation, ponctuation, détection d'entités nommées, ...) ont été identifiées. Les formes erronées et leurs phrases associées qui résulteraient de telles erreurs sont donc exclues de celles qui nous intéressent ici.

prétations, ou lorsque les ressources utilisées (lexique, grammaire ...) ne sont pas assez restrictives et acceptent un langage plus large.

Afin de rejeter les phrases n'appartenant pas à la langue, on souhaite disposer d'un lexique le plus précis et détaillé possible. En effet, plus une forme lexicale est spécifiée, moins elle se combine avec les autres constituants de la phrase, et par conséquent, moins elle permet d'interprétations incorrectes.

2.1 Causes d'échec d'une analyse

Chaque forme possède, à travers ses lemmes, différentes informations pouvant être regroupées en deux ensembles : d'une part la catégorie syntaxique (nom, verbe, adjectif, ...), d'autre part les informations morphologiques (nombre, genre, personne, temps, mode, ...) et syntaxiques (valence, facultativité des arguments, réflexivité, passivation, ...). L'échec d'une analyse à cause d'une forme est la conséquence d'un problème touchant à au moins un de ces ensembles.

2.1.1 Défaut de catégorisation

Une forme peut être associée à plusieurs lemmes (homonymes) avec des catégories syntaxiques distinctes. Le traitement de telles formes ambiguës au sein d'une phrase se gère par le passage d'un treillis de mots (ou DAG) à l'analyseur syntaxique (Sagot & Boullier, 2005). Une analyse syntaxique réussie valide au moins un chemin possible de lecture dans ce treillis.

Cependant, un lexique peut ne pas recenser tous les homonymes d'une forme et induire ainsi des échecs d'analyse. Par exemple, la forme « fiche » dénote un nom commun et une flexion du verbe « ficher » . S'il n'existe aucun lemme associé de catégorie *nom-commun*, la phrase « Ma fiche contient une erreur. » sera représentée par une seule séquence de catégories *ma/pronom-possessif fiche/verbe contient/verbe une/det erreur/nom-commun*. À moins qu'une production grammaticale n'accepte une telle construction, son analyse devrait aboutir à un échec.

2.1.2 Sur-spécification

En général, on associe aux règles de grammaire des décorations, exprimées sous formes de structures de traits et chargées de compléter les vérifications amorcées par le squelette syntaxique d'une production grammaticale (Abeillé, 1993). Par exemple, un squelette vérifie la présence d'un groupe nominal sujet et d'un verbe dans une phrase là ou les décorations en vérifient l'accord (même personne, nombre, et éventuellement genre).

Comme nous l'avons expliqué, il est souhaitable que les formes lexicales soient les plus spécifiées possible afin de réduire les ambiguïtés. En revanche, si ces dernières sont trop restrictives (autrement dit sur-spécifiées), certaines analyses échouent à cause du mécanisme d'unification des décorations de la grammaire et des restrictions d'utilisation des entrées lexicales.

Il est par exemple très difficile de renseigner un verbe sur l'ensemble de ses emplois possibles, du fait de la polysémie, de la facultativité de certains arguments, de possibles alternations (« acheter qchose » donnant « qchose s'achète »), et de multiples réalisations des arguments (« aimer qchose » , « aimer que + S » , « aimer Sinf »). Il arrive donc que l'on considère comme obligatoires des aspects qui ne sont que facultatifs dans certains cas. Ce constat s'étend aux autres catégories syntaxiques dès lors qu'on leur attache des cadres de catégorisation.

2.2 Réanalyser en sous-spécifiant

Puisque seules les phrases dont l'analyse a échoué sont conservées durant l'étape de fouille d'erreurs, leur taux d'analyse est nul. Si une modification des informations lexicales portées par une forme suspecte f permet d'augmenter sensiblement le taux de réanalyse des phrases qui lui sont associées, il est raisonnable de penser que le problème est bien lié à f. La difficulté est alors de trouver quelles modifications permettent des augmentations sensibles. Plutôt que de tester toutes les combinaisons de modifications possibles, ce qui est exponentiel, nous nous reposons sur la capacité de notre analyseur à pouvoir gérer des formes sous-spécifiées.

Une fois obtenus de nouveaux résultats d'analyse, nous sommes en mesure d'en extraire des hypothèses de correction (voir Sect. 2.2.2).

2.2.1 Génération et utilisation de jokers

Afin de rendre analysables des phrases qui ne l'étaient initialement pas, nous introduisons à la place des formes lexicales suspectes des formes sous-spécifiées appelées *jokers*. Dans l'approche actuelle (qui demande à être affinée), elles ne possèdent qu'une catégorie syntaxique (parmi les catégories « ouvertes » : verbe, nom commun, adjectif ou adverbe). Elles n'ont donc aucune information morphologique ou syntaxique fixe et remplissent toujours les conditions fixées par les décorations des productions grammaticales. Puisque leur utilisation ne soulève aucun conflit lors des analyses (excepté pour la catégorie syntaxique), les substituer à une forme suspecte dans une phrase rejetée favorise la réussite de son analyse. Cependant, cela peut introduire une certaine ambiguïté car il n'y a plus de filtrage au niveau des décorations.

Étant donné que nous ne pouvons savoir *a priori* quel type d'erreur (sur-spécification ou défaut de catégorisation) est responsable des échecs d'analyse, nous considérons les deux simultanément au moment de générer les jokers.

Pour envisager une sur-spécification, nous remplaçons une forme de catégorie X par un joker de même catégorie X. Les caractéristiques permettent alors d'explorer les mêmes productions grammaticales que pour la phrase initiale, sans pour autant être arrêté par les décorations.

Pour faire face à un défaut de catégorisation d'une forme f, nous créons des jokers avec des catégories syntaxiques différentes de celles initialement recensées pour f. En procédant ainsi, les réanalyses exploreront d'autres productions. Ces jokers sont générés à partir des informations fournies par un lemmatiseur (*stemmer*) ou par un tagger probabiliste tel que TREETAGGER (Schmid, 1999).

Nous aurions pu utiliser un joker unique ne possédant même pas de catégorie syntaxique et permettant de couvrir à lui seul l'ensemble des situations décrites ci-dessus. Cependant, un tel joker introduit une très forte ambiguïté, aboutissant soit à un échec des analyses par limite de temps ou de mémoire, soit à la surgénération de représentations pour une phrase. Dans le premier cas, nous ne collectons aucune donnée, dans le second cas, le volume de données est trop important pour être correctement trié et valorisé. Notre approche permet (en grande partie) d'écarter ces problèmes tout en évitant de multiplier le nombre de jokers par forme suspecte.

Nous avons testé une moyenne de 2.05 jokers par forme suspecte (10978 au total), donnant lieu à 117655 nouvelles analyses.

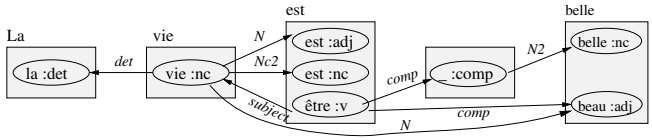

FIG. 1 – Extrait de la représentation graphique d'une forêt partagées de dépendances

2.2.2 Extraction de signatures syntaxiques

Si une forme suspecte a été correctement identifiée, son remplacement par des jokers dans les phrases qui lui sont associées permet à certaines analyses de réussir. Dans les faits, on observe une relation nette entre le taux de succès de l'analyse des phrases modifiées et le taux de suspicion de la forme concernée.

Notre analyseur renvoie l'ensemble des interprétations possibles d'une phrase sous la forme d'une forêt partagée de dépendances (Fig. 1) où les nœuds représentent les lemmes et les arcs les dépendances syntaxiques entre les lemmes. Chaque nœud possède des informations relatives au lemme et à la production grammaticale ancrée (dans le cadre d'une grammaire lexicalisée). Chaque dépendance est caractérisée par un nœud gouverneur source, un nœud gouverné cible, une nature et un label qui dépend de la grammaire. Ce label dénote souvent (mais malheureusement pas toujours) la fonction syntaxique de la cible (sujet, objet, ...). Afin de gérer les ambiguïtés, des informations complémentaires locales au nœud gouverneur lient les lemmes et les dépendances à une ou plusieurs interprétations. Ainsi, la représentation issue de l'analyse de la phrase pour « La vie est belle » (Fig. 1) donne lieu à quatre lectures possibles, du fait (a) de l'ambiguïté de « est » comme verbe à copule, nom commun (en apposition de « vie ») et adjectif ainsi que (b) de l'ambiguïté de « belle » entre adjectif et nom.

Sans aucune information supplémentaire, les deux interprétations comme nom et adjectif de « est » auraient dû être rejetées car introduisant une apposition rare et/ou construisant une phrase sans verbe.

Les forêts contiennent donc les dépendances entrantes et sortantes depuis et vers un joker. Nous appelons désormais *signature syntaxique* l'ensemble de dépendances autour d'un joker dans une interprétation particulière et *groupe de signatures* l'ensemble des signatures syntaxiques possibles extraites des interprétations obtenues par l'analyse réussie d'une phrase.

Ces signatures représentent les conditions dans lesquelles l'analyse a pu aboutir, i.e. les données que la grammaire aurait accepté pour la forme suspecte. Du fait de l'ambiguïté consécutive à l'introduction d'un joker, un analyseur peut produire plusieurs interprétations et donc plusieurs signatures. Parmi ces interprétations, une est plus proche du sens réel de la phrase que les autres. La signature qu'elle contient possède alors les données les plus pertinentes et intéressantes, celles que nous recherchons afin de déterminer les corrections à appliquer au lexique.

3 Identifier les meilleures signatures

En se plaçant au niveau d'un seul groupe de signatures (produit à partir d'une seule phrase), nous sommes incapables de différencier les signatures pertinentes de celles qui ne sont qu'une

conséquence de l'ambiguïté introduite par le joker.

La variabilité de contexte induite par plusieurs groupes de signatures (produits à partir de plusieurs phrases) nous apporte une solution à ce problème. En effet, elle implique la diversification des signatures « parasites » qui contraste avec la stabilité des signatures pertinentes représentant le(s) sens réel(s) de la forme.

Une répétition bien marquée de certaines signatures sur l'ensemble des phrases suggère alors un schéma d'utilisation attendu par la grammaire pour la forme. Afin de pouvoir l'observer, nous valorisons/dévalorisons les signatures par le biais d'un calcul statistique simple en deux étapes.

Première étape : distribution locale des poids entre signatures. L'intérêt que nous portons à un groupe de signatures dépend de sa taille : plus il contient de signatures moins il présente d'intérêt. En effet, il est vraisemblable que plusieurs squelettes syntaxiques « permissifs » lui correspondent, à l'image de ceux permettant les diverses interprétations illustrées par la figure 1. Pour chaque groupe g, nous calculons donc un poids $P = c^n$ avec c une constante incluse dans $]0, 1[$ (par exemple $0,95$) et n la taille du groupe.

Au niveau d'un groupe, toutes les signatures sont d'égale importance, nous répartissons donc de manière équitable les poids attribués au groupe : chacune signature reçoit un poids $p_g = \frac{P}{n} = \frac{c^n}{n}$ qui dépend donc doublement de la taille du groupe.

Seconde étape : calcul global des poids. Une fois l'étape précédente réalisée, nous additionnons les poids obtenus par une même signature σ dans les différents groupes où elle apparaît pour calculer son score $s_\sigma = \Sigma_g p_g$.

Les meilleures signatures, à savoir celles qui se trouvent dans plusieurs groupes et dans des groupes de petite taille, reçoivent alors un score s_σ plus élevé.

4 Synchronisation lexique-grammaire

Cette technique permet à une grammaire d'exprimer ses attentes pour les formes suspectes. Si elle n'est pas parfaite, les représentations qu'elle produit ainsi que les signatures que l'on en extrait ne le sont pas non plus. En fait, dans le cas où la grammaire est parfaite, nous pouvons qualifier les suggestions faites par cette technique comme permettant une correction du lexique. Dans le cas inverse, il s'agit alors d'une technique permettant de diminuer le nombre de conflits entre une grammaire et un lexique, i.e. permettant une meilleure « synchronisation » entre le lexique et la grammaire.

Il est à noter qu'un ensemble de signatures incorrectes représente une source d'informations intéressante sur les manques et incorrections d'une grammaire.

5 Résultats

Le travail présenté ici, tout comme la technique de fouille d'erreurs, est un mécanisme de retour sur erreurs. Ce terme désigne des mécanismes réutilisant les erreurs produites par un programme afin d'améliorer automatiquement ou semi-automatiquement sa qualité. De manière à garantir que l'origine des erreurs produites est effectivement le programme, les données ana-

lysées doivent être fiables. Dans le cas présent, les erreurs sur lesquelles nous travaillons sont issues d'une campagne d'analyse d'un corpus MD de 331 000 phrases extraites du *Monde diplomatique* réalisée durant la validation de la technique de fouille d'erreurs (Sagot & Villemonte de La Clergerie, 2006).

Le lexique que nous cherchons à améliorer est le *Lefff* (***Le**xique des **f**ormes **f**léchies du **f**rançais*) (Sagot *et al.*, 2006). En partie acquis automatiquement, ce lexique morpho-syntaxique à large couverture du français est en constant développement et possède, à l'heure actuelle, plus de 520 000 entrées. La grammaire FRMG (Thomasset & Villemonte de La Clergerie, 2005) que nous utilisons est une grammaire hybride TAG/TIG avec décorations. Elle est construite à partir d'une *méta-grammaire* plus abstraite qui produit un ensemble de 134 arbres très factorisés. Malgré son très faible nombre d'arbres, sa factorisation lui permet de couvrir un grand nombre de cadres de catégorisation pour les verbes, la passivation, les extractions (relatives, interrogatives, clivées), certaines inversions du sujet, certaines constructions à verbe support (« faire attention à »). Néanmoins, nombre de phénomènes ne sont pas encore traités (comme la sous-catégorisation sur les adjectifs et les noms). La grammaire FRMG couplée à *Lefff* assurait en 2005 une couverture de l'ordre de 41% sur le corpus MD.

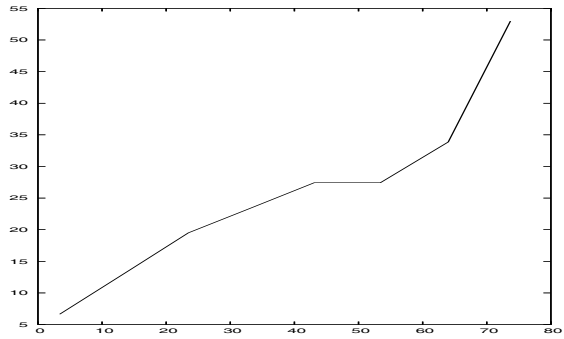

FIG. 2 – Taux de réussite des réanalyses (axe Y) en fonction des taux de suspicion (axe X)

5.1 Exactitude de la détection automatique des formes suspectes

La courbe de la figure 2 nous permet d'observer une corrélation très nette entre les taux de réussite des réanalyses et les taux de suspicion des formes. Cela atteste la validité des informations produites par l'étape précédente de fouille d'erreurs.

Les valeurs présentées par cette courbe sont en réalité des moyennes calculées après un regroupement des formes suspectes par intervalle de taux de suspicion. Sans cela, la courbe présente des variations rendant difficile son observation.
Ces variations s'expliquent principalement par le fait que certaines formes ont été suspectées à la place de la grammaire ce qui explique que leur échange avec des jokers n'ait rien apporté. En effet, certaines formes ont une affinité marquée pour des constructions spécifiques ; par exemple une inversion du sujet en présence de l'adjectif 'rare' comme dans « Rares sont ceux qui tentent

d'en sortir. » ou 'nombreux' dans « Nombreux sont ceux qui refusent. » [2]. Ces formes ont alors payé cette affinité par une suspicion injustement élevée à leur égard.
Une autre raison moins importante expliquant ces variations est que l'utilisation de jokers augmente sensiblement le taux de *timeout* pour les phrases, et cela même en ayant imposé une limite de 40 mots sur la longueur des phrases analysées.

Puisque ces deux phénomènes s'observent à tous les niveaux de taux de suspicion, le regroupement des valeurs par intervalle a permis d'en diminuer l'influence sur la courbe de la figure 2.

Toujours dans une optique de retour sur erreurs, notons qu'il est tentant de voir les phrases des suspects forts avec de faibles taux de réanalyse comme indiquant des manques de la grammaire. Il serait alors intéressant de les analyser au moyen d'un système d'inférence grammaticale.

5.2 Évaluation de la qualité des signatures

Afin d'évaluer la qualité des signatures produites, nous avons ordonné les formes suspectes en accord avec le calcul suivant : $M_f = S_f.ln(NS_f)$, S_f étant le taux de suspicion d'une forme et NS_f le nombre de phrases associées[3]. Nous avons ensuite examiné nombre d'entre elles à travers une interface Web (Fig. 3) nous permettant d'accéder, pour chaque forme, aux jokers testés, aux taux de réanalyses obtenus, aux phrases testées et aux meilleures signatures retenues. De même, elle nous permet de laisser des commentaires et de soumettre des requêtes au lexique et à l'analyseur. À terme, cette interface a vocation à être utilisée par des linguistes.

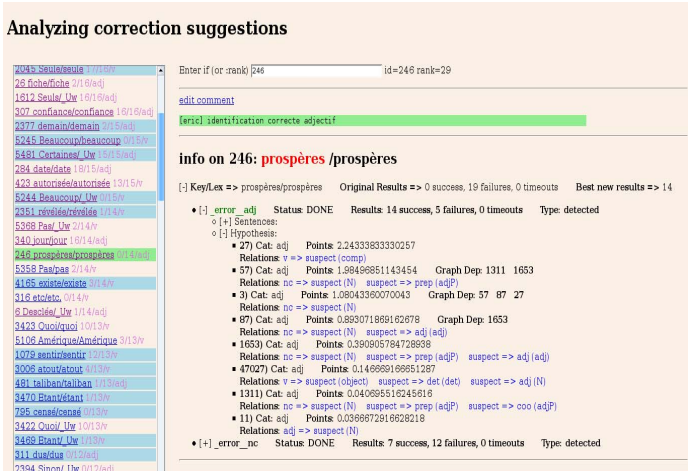

FIG. 3 – Interface d'exploration des signatures

Lors de l'étude des meilleures signatures, certains doutes ont été confirmés : notre technique manque de maturité. Nous avons identifié un certain nombre de phénomènes nous empêchant de correctement quantifier la qualité des signatures. Cependant, nous savons déjà comment faire face à la plupart (voir Sect 6).

[2] Ces exemples reflètent aussi le style recherché du corpus journalistique étudié !
[3] Un fort taux de réanalyse sur un nombre réduit de phrases est peu significatif.

Toutefois, dans bien des cas, nous avons obtenu des résultats pertinents et instructifs qui nous ont permis d'améliorer nos outils (et pas seulement notre lexique). Par exemple, on retrouve la bonne signature comme dans le cas de « prospères » où l'on retrouve un usage d'adjectif épithète (joker + signature), alors qu'il n'existe que comme verbe dans *Lefff*. Pour la forme verbale « révéler » , les hypothèses font ressortir qu'elle attend bien un argument attributif (« ce choix pourrait se révéler catastrophique. ») mais qu'il lui manque le côté réflexif, à cause de constructions prépositionnelles comme « contraint de révéler X » ou « penser à révéler X » .

Bien que devant encore mûrir, notre approche s'est montrée viable. Nous continuerons à la développer afin d'obtenir un outil pleinement fonctionnel. L'achèvement de certaines améliorations donnera notamment lieu à de nouvelles campagnes de calcul.

6 Développements futurs

Durant nos expériences, nous avons pu établir une liste de problèmes à traiter et des solutions pour les résoudre :

- Il est très fréquent de pouvoir appliquer plusieurs productions grammaticales à une suite de formes, surtout si la catégorie syntaxique d'une de ces formes varie (comme pour les jokers). Cependant, ces productions n'ont pas les mêmes fréquences d'utilisations et par conséquent, les signatures qui en résultent ne représentent pas la même quantité d'information utile. De telles données sur les fréquences d'utilisation nous seraient utiles afin de pondérer les signatures et de diminuer l'ingérence de signatures « parasites » dans les résultats.
- Les signatures doivent être nettoyées pour éliminer l'adjonction de certains adjoints (gouvernés par les suspects) qui ne sont pas primordiaux pour caractériser ceux-ci. Cela nous permettrait de consolider des signatures actuellement séparées par des adjoints inutiles. Néanmoins, à ce stade, il n'est pas toujours évident de juger de l'importance d'un adjoint.
- Il nous faut regrouper les formes par famille de lemmes sous-jacents de manière à augmenter la variabilité des contextes testés et ainsi cerner ce qu'ils ont en commun. Néanmoins, il faut garder à l'esprit que certains problèmes ne se manifestent que pour quelques formes, par exemple une mauvaise attribution de l'auxiliaire à utiliser pour des participes passés (exemple de « larvé » faussement listé dans *Lefff* comme utilisant l'auxiliaire « avoir »). Nous avons aussi mentionné que, parfois, le problème résulte du manque dans le lexique d'un des lemmes possibles pour une forme suspecte.
- Il nous faut regrouper les signatures qui traduisent en fait un même phénomène syntaxique sous des aspects différents ; comme par exemple : le sujet et autres arguments verbaux ont diverses réalisations (nominales, cliticisées, pronoms relatifs, pronoms interrogatifs), ou encore un verbe avec objet sous forme active et passive. Le regroupement des formes par lemme est susceptible d'aider.
- Certaines formes suspectes donnent des signatures équivalentes aux informations syntaxiques déjà présentes dans le lexique. Ce genre de cas implique que les signatures sont incomplètes. À l'heure actuelle, elles manquent principalement d'informations morphologiques. L'intégration de ces informations déjà présentes dans les forêts de dépendances, mais non encore exploitées, représente donc la prochaine étape dans l'amélioration du modèle des signatures.
- Certaines formes ont été injustement suspectées à cause de leur affinité avec des constructions syntaxiques non gérées par la grammaire. L'utilisation de plusieurs analyseurs syntaxiques avec des grammaires différentes durant l'étape préalable de fouille d'erreurs permettraient éventuellement de filtrer une partie des formes suspectes.

– Les signatures sont composées d'un ensemble de dépendances syntaxiques entre les mots et le joker dans les représentations générées d'une phrase. Ces signatures dépendent directement de la grammaire utilisée et peuvent être difficiles à comprendre pour une personne non familière avec ce formalisme. Un effort doit donc être réalisé pour les traduire vers une représentation indépendante de la grammaire et plus humainement compréhensible.

7 Conclusion

Les expériences présentées confirment en premier lieu la capacité de la technique de fouille d'erreurs à identifier de bonnes formes suspectes. Leur transformation en jokers augmente le taux de réanalyses réussies de manière coordonnée avec le taux de suspicion d'une forme.

En second lieu, elles valident la faisabilité d'un mécanisme automatique de suggestion de corrections lexicales sur les formes suspectes (i.e. sur les lemmes sous-jacents). Elles montrent qu'il est également possible d'obtenir du retour d'information sur des manques grammaticaux.

Néanmoins, un travail reste encore à faire pour affiner la qualité des corrections suggérées en distinguant mieux l'essentiel de l'accessoire dans les signatures, notamment à travers des améliorations introduites précédemment.

Références

ABEILLÉ A. (1993). *Les nouvelles syntaxes, grammaire d'unification et analyse du français*. Armand Colin.

BRENT M. R. (1993). From grammar to lexicon : unsupervised learning of lexical syntax. *Computational Linguistic*, **19**(2), 243–262.

CRYSMANN B., FRANK A., KIEFER B., KRIEGER H.-U., MÜLLER S., NEUMANN G., PISKORSKI J., SCHÄFER U., SIEGEL M., USZKOREIT H. & XU F. (2002). An integrated architecture for shallow and deep processing. In *Proceedings of the 40th Annual Meeting of the ACL*, p. 441–448.

GROVER C. & LASCARIDES A. (2001). XML-based data preparation for robust deep parsing. In *Meeting of the Association for Computational Linguistics*, p. 252–259.

HORIGUCHI K., TORISAWA K. & TSUJII J. (1995). Automatic acquisition of content words using an HPSG-based parser. In *Proceedings of NLPRS'95.*

SAGOT B. & BOULLIER P. (2005). From raw corpus to word lattices : robust pre-parsing processing. In *Proceedings of L&TC*, Poznan, Pologne.

SAGOT B., CLÉMENT L., VILLEMONTE DE LA CLERGERIE É. & BOULLIER P. (2006). The Lefff 2 syntactic lexicon for french : architecture, acquisition, use. In *Proceedings of LREC'06*.

SAGOT B. & VILLEMONTE DE LA CLERGERIE É. (2006). Trouver le coupable : Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. In *Proceedings of TALN'06*, p. 287–296.

SCHMID H. (1999). Probabilistic part-of-speech tagging using decision trees. *IMS-CL.*

THOMASSET F. & VILLEMONTE DE LA CLERGERIE É. (2005). Comment obtenir plus des méta-grammaires. In *Proceedings of TALN'05*, Dourdan, France : ATALA.

VAN NOORD G. (2004). Error mining for wide-coverage grammar engineering. In *Proceedings of ACL 2004*, Barcelone, Espagne.