Régler les règles d'analyse morphologique

Bruno CARTONI
ISSCO/TIM/ETI – Université de Genève
40 bd du Pont d'Arve, 1205 Genève
bruno.cartoni@eti.unige.ch

Résumé. Dans cet article, nous présentons différentes contraintes mécaniques et linguistiques applicables à des règles d'analyse des mots inconnus afin d'améliorer la performance d'un analyseur morphologique de l'italien. Pour mesurer l'impact de ces contraintes, nous présentons les résultats d'une évaluation de chaque contrainte qui prend en compte les gains et les pertes qu'elle engendre. Nous discutons ainsi de la nécessaire évaluation de chaque réglage apporté aux règles afin d'en déterminer la pertinence.

Abstract. In this article, we present various constraints, mechanical and linguistic, that can be applied to analysing rules for unknown words in order to improve the performance of a morphological analyser for Italian. To measure the impact of these constraints, we present an evaluation for each constraint, taking into account the gains and losses which they generate. We then discuss the need to evaluate any fine-tuning of these kinds of rules in order to decide whether they are appropriate or not.

Mots-clés : évaluation, analyse morphologique, mots inconnus, morphologie constructionnelle.
Keywords: evaluation, morphological analysis, unknown words, constructional morphology.

1 Introduction

Les mots inconnus représentent un problème récurrent pour le traitement automatique de la langue. Traditionnellement, on distingue trois types de mots inconnus : les noms propres, les erreurs et les mots issus de la créativité lexicale (néologismes), chaque type recevant un traitement particulier. Pour les mots inconnus issus de la créativité lexicale et particulièrement de la créativité morphologique (les néologismes construits), il a souvent été proposé d'employer des règles d'analyse qui formalisent plus ou moins les procédés de construction des mots. Ces règles, souvent basées sur des principes linguistiques, mais également fortement contraintes par les ressources disponibles et d'autres considérations plus mécaniques, rencontrent un certain nombre de limites. En effet, ce type d'approche est rapidement confronté au problème de l'homographie des chaînes de caractères des mots potentiellement construits avec des mots qui ne le sont pas. Ainsi, il arrive que ces règles d'analyse engendrent plus de problèmes qu'elles n'en résolvent. Pour pallier ces écueils, ces

règles sont ajustées au moyen d'un certain nombre de contraintes, linguistiques ou mécaniques.

Dans cet article, nous présentons l'évaluation des contraintes linguistiques et mécaniques qui peuvent être appliquées sur les règles d'analyse des mots inconnus. Nous commençons par passer en revue quelques analyseurs à base de règles qui traitent des mots inconnus morphologiquement construits (section 2), puis nous présentons notre analyseur basé sur les chaînes de caractères et sur des contraintes qui permettent d'éviter les écueils (section 3). Le nombre de mauvaises analyses engendrées par les règles ou par l'application d'une contrainte linguistique peut cependant ternir le bénéfice apporté par les règles dans le traitement des mots inconnus. L'évaluation de notre analyseur et de chacune des contraintes représente par conséquent un sujet crucial. Après avoir posé les jalons de l'évaluation de l'analyse morphologique (section 4.1), nous présentons les résultats de l'évaluation des contraintes de notre analyseur (section 4.2), afin de voir quelles sont celles qui aident le système à tendre vers une performance maximale. Nous concluons en montrant que l'importance donnée aux pertes et aux gains induits par ces contraintes dépend avant tout de la finalité de la tâche.

2 Analyse morphologique et mots inconnus : état de l'art

Les mots inconnus sont un phénomène constant, mais leur proportion relativement restreinte constitue un frein aux méthodes d'apprentissage automatique et nous pousse à privilégier des méthodes davantage linguistiques, à base de règles.

Une grande partie des études qui exploitent les connaissances morphologiques pour traiter les mots inconnus néologiques se concentrent sur l'incomplétude lexicale des lexiques des analyseurs morphosyntaxiques. Elles ont pour principal objectif de deviner la catégorie morphosyntaxique des mots inconnus, le plus souvent en exploitant les terminaisons typiques de ces mots (Guilbaud et al., 1997) et (Woods, 2000). Si certaines études ne se réclament d'aucune approche linguistique particulière, d'autres montrent explicitement leur référence théorique, à l'image de (Byrd, 1983) et (Byrd et al., 1989), qui proposent une application des principes de la morphologie lexématique de (Aronoff, 1976). (Byrd, 1983) prône un véritable calcul morphologique permettant de retrouver la base et le(s) affixe(s) qui constituent le mot construit.

Toutes ces études prévoient d'ajuster les règles par l'entremise de contraintes qui permettent d'éviter les problèmes engendrés par ce type d'approche (cf. ci-après section 3.1). Une des contraintes principales porte sur la catégorie morphosyntaxique de la base, qui peut facilement être filtrée, étant donné qu'elle se trouve dans le lexique. D'autres réglages sont davantage sémantiques, comme l'étymologie latine de certains mots anglais qui favorise l'affixation avec un affixe latin (Byrd et al., 1989).

Du point de vue de l'évaluation, les études mentionnées ne font état ni des gains apportés par ces règles ni de la correction des analyses, et se cantonnent à évaluer d'un point de vue beaucoup plus large l'amélioration des performances de l'analyseur en général. Or, nous pensons que le traitement des mots inconnus est, de par sa nature, une tâche extrêmement précise et granulaire. Une augmentation même minime de la performance de l'analyseur doit être mise en regard non seulement des mauvaises analyses engendrées (le bruit), mais également du nombre de mots non-analysés (le silence). En effet, si une règle d'analyse engendre plus de problèmes qu'elle n'en résout, le gain global est alors insignifiant. Dans la suite, nous présentons les différentes questions soulevées par l'évaluation de l'analyse

morphologique à base de règles. Mais avant cela, nous présentons l'analyseur de mots inconnus préfixés que nous avons contraint puis évalué.

3 Analyseur morphologique basé sur les règles

Dans le cadre d'un projet de recherche plus large de traduction automatique des mots inconnus morphologiquement construits (Cartoni, 2005), nous avons mis au point un analyseur morphologique de l'italien permettant d'analyser les mots inconnus mais corrects construits par préfixation. Par analyse morphologique, nous entendons l'identification des mots réellement préfixés et donc l'individualisation de leur base. Le calcul sémantique base+préfixe est effectué ultérieurement par les règles. Cet analyseur s'appuie sur un lexique de référence de l'italien (*Mmorph* (Petitpierre et al., 1995) – 739 000 formes) dont il extrait les informations lexicales. L'analyseur morphologique est donc constitué de *règles de construction des mots* (RCM[1]), comme le montre la figure 1 :

```
(1)   RCM (iper) :
(2)   X = iper_PREF [Y]
(3)   Y ∈ L_IT
```

Figure 1 : Règle d'analyse pour la préfixation en *iper*

La règle de la figure 1 analyse les mots inconnus dont la première séquence de lettres est `i p e r`. Si la séquence de lettres restantes (c'est-à-dire la base potentielle) est présente dans le lexique de référence (ligne (3)), le mot X est alors considéré comme construit. Toutes les recherches que nous avons déjà évoquées (section 2) prenaient évidemment en compte le fait qu'un mot construit l'était avec un affixe (instancié par la règle) et une base qui devait être connue du lexique de référence. Il nous faut toutefois mentionner que cette évidence ne s'applique pas toujours, notamment lorsqu'un préfixe s'accole avec un trait d'union à une base absente du lexique, formant tout de même un mot construit (il s'agit souvent de noms propres, comme *pro-Tibet*). Ajoutons également que, comme les règles traitent des chaînes de caractères, nous devons formaliser plusieurs règles pour un même préfixe en fonction des différents allomorphes qu'il peut avoir (le préfixe *in* peut par exemple prendre la forme *il*, *im*, ou *ir,* suivant la consonne initiale de la base), ou encore avec ou sans trait d'union, l'usage de celui-ci étant passablement flottant.

Une étude de faisabilité (Cartoni, 2006) nous avait déjà montré que, parmi les 46 préfixes productifs de l'italien listés par (Iacobini, 2004), certains sont très peu problématiques en terme de transparence et d'homographie avec d'autres chaînes de caractères. Nous les avons donc pour l'instant mis de côté et nous nous sommes contentés d'implémenter l'analyseur avec les 14 préfixes de l'italien (et leurs allomorphes) qui posent le plus de problèmes, à savoir : *pro, dis, trans, inter, in, poli, arci, retro, post, mini, iper, multi, ri* et *co*.

[1] La morphologie lexématique rejette la notion de mot et préfère parler de lexèmes, comme unité abstraite. Il en résulte que l'on parle maintenant de *règles de construction des lexèmes* (RCL) plutôt que de RCM. Nous conservons cependant l'appellation de RCM, étant donné que, d'un point de vue informatique, il s'agit bien de mots (dans le sens de mot forme) que nous voulons analyser.

3.1 Les problèmes de l'analyse morphologique

De telles règles, bien que très simples, se révèlent relativement efficaces, mais elles présentent évidemment un certain nombre d'écueils, à cause de l'homographie de certaines chaînes de caractères avec des mots réellement construits. (Grabar et al., 2006) distinguent cinq types de mauvaises analyses engendrées par ce genre de méthode : (1) des « lexèmes dans lesquels l'opération étudiée n'est pas la dernière opération constructionnelle », (2) des « lexèmes difficilement analysables comme construits en français », (3) des « lexèmes comportant une suite graphique accidentellement identique aux affixes étudiés », (4) des « lexèmes polysémiques [dont] le sens attesté n'est pas celui qui nous intéresse » et (5) des « erreurs et fautes d'orthographe ».

Dans la présente étude, nous appliquons les règles aux mots inconnus du lexique de référence. Par conséquent, nous rencontrons majoritairement des problèmes du type (1), quand l'opération qui a construit le mot inconnu n'est pas la préfixation (`"prostatiche" = prostata +ico,` et non pas `pro+statiche`), ainsi que des problèmes du type (5) où le mot inconnu est erroné, mais a été analysé comme une séquence préfixe + base (`"progesso"` est la forme erronée de `"progresso"` et non pas `"pro+gesso"`). Il reste néanmoins quelques cas de mauvaises analyses qui proviennent de l'absence du mot dans le lexique de référence bien qu'il ne s'agisse ni d'un néologisme construit ni d'une faute d'orthographe. Il s'agit alors majoritairement d'emprunts ou de termes techniques.

Tout l'enjeu de l'analyseur est par conséquent d'éviter les pièges provoqués par ces homographies. Ainsi, nous avons *réglé nos règles* avec un certain nombre de contraintes que nous décrivons ci-dessous, et que nous évaluons dans la section 4.

3.2 Les contraintes possibles sur les règles

Pour *régler nos règles*, nous avons mis en place deux types de contraintes qui étaient déjà proposées notamment par (Byrd et al., 1986 et Bopp et al., 2004). La première, qui prend en compte la catégorie morphosyntaxique de la base, est basée sur des principes linguistiques. La seconde, qui porte sur la valeur sémantique de la base, est également d'inspiration linguistique, même si certaines données proviennent d'intuitions plus empiriques.

La contrainte de la catégorie morphosyntaxique de la base est motivée par le fait que certains affixes ne s'accolent qu'à certains types de base, même s'il est vrai que plus d'un quart des préfixes de l'italien s'accolent aux trois catégories lexicales majeures (adjectif, nom, verbe) (Iacobini, 2004). Ainsi, pour la règle de préfixation en *mini,* nous avons contraint la règle avec une catégorie de base uniquement nominale. Pour les règles de préfixation en *pro, retro, post, poli, multi, trans, arci,* et *iper*, la base doit être soit adjectivale, soit nominale. Enfin, pour les bases des règles de préfixation en *dis, ri, co,* et *inter,* les trois catégories majeures sont possibles. Notons également que cette contrainte, même si elle est parfois très large, permet d'exclure des mauvaises analyses sur un déterminant ou une conjonction (`arcipel` est un emprunt, et non pas une construction avec le préfixe *arci* et la préposition contractée `pel`).

La deuxième contrainte porte sur la valeur sémantique de la base qui peut favoriser également l'application de tel ou tel préfixe. A moins de disposer de ressources lexicales contenant des informations sémantiques, il est très difficile de formaliser informatiquement ce genre de contrainte. Cependant, la sémantique de la base est parfois observable dans sa forme de

surface. Ainsi, comme l'affirme (Iacobini, 2004, p. 114) « l'emploi d'un préfixe peut être conditionné par le suffixe de la base », étant donné que la valeur sémantique de la base peut être exprimée par l'emploi d'un ou plusieurs suffixes. Cette assertion est intéressante car le ou les suffixe(s) concernés peuvent être exprimés en termes de chaînes de caractères, permettant de contraindre la règle d'analyse. Cette optique revient à dire que certains préfixes sont productifs sur des bases qui sont déjà des mots construits, ou que la « constructivité » des bases permet la préfixation. (Krott et al., 1999) et plus tard (Namer, 2003) ont souligné le nombre important de mots construits sur des bases elles-mêmes construites. Il est donc envisageable, pour certains préfixes, de contraindre la base sur certaines terminaisons que nous considérons alors comme des « indices de constructivité ».

Dans la mesure où certains préfixes comme *ri, co* et *retro* sont réputés productifs avec les noms déverbaux, nous pouvons contraindre les règles d'analyse en imposant la présence de suffixes typiques de la nominalisation déverbalisante, comme *–zione* et *–mento*. Il en va de même pour le préfixe *co* qui est très productif sur des noms d'agent (typiquement suffixés en *–(t)ore*). (Iacobini, 2004) cite également les adjectifs en *-bile* qui sont fréquemment préfixés en *in*. Le corollaire de cette dernière remarque est qu'un grand nombre de noms en *-ità* (suffixation nominale des adjectifs en *-bile*) sont également préfixés en *in* (comme *inconciliabilità, indisponibilità*).

De plus, la préfixation permet la formation d'adjectifs sur des bases nominales, qui prennent la forme de l'adjectif relationnel correspondant. En français, par exemple, *anticancéreux* est formé sur *cancéreux,* qui est l'adjectif relationnel de *cancer*. Même si sémantiquement la base du mot construit est le nom qui est à la base de l'adjectif (*anticancéreux = contre le cancer* et non pas *contre les cancéreux*), la forte régularité de ce type de construction nous permet de considérer la base comme un adjectif relationnel (indépendamment du calcul sémantique nécessaire à l'interprétation du mot construit). Les adjectifs relationnels sont eux aussi des mots construits sur des bases nominales, à l'aide de suffixes typiques de ce genre de formation. Il est donc intéressant de contraindre les bases analysées en fonction des suffixes types de la formation d'adjectifs relationnels. Pour l'italien, les suffixes typiques de formation des adjectifs relationnels sont : *-ale, -are, -ario, -ano, -ico, -ile, -ino, -ivo, -orio, -esco, -asco, -iero, -izio, -aceo* (Wandruszka, 2004). Comme nous travaillons sur des chaînes de caractères, il faut évidemment décupler ces suffixes en fonction de chaque flexion de genre et de nombre (*-ino, -ina, -ini, -ine*).

Ainsi, pour certains préfixes implémentés jusqu'à présent dans notre analyseur, nous avons pu contraindre les règles avec les indices suivants : (a) les indices d'adjectif relationnel pour les préfixes *inter, multi, poli, post,* et *trans ;* (b) les indices de noms d'action (*-zione, -mento*) pour les préfixes *co, retro* et *ri ;* (c) les indices de noms d'agent (*-(t)ore*) pour le préfixe *co,* et enfin, (d) les indices d'adjectifs en *-bile* et de noms en *-ità* pour le préfixe *in.* Evidemment, la validité de ces contraintes doit être vérifiée sur une large échelle, vérification que nous présentons ci-dessous.

4 Évaluation de l'analyse morphologique des mots inconnus

Nous l'avons dit, la plupart des recherches qui exploitent les propriétés morphologiques pour résoudre le problème des mots inconnus construits n'évaluent que le produit final (la couverture lexicale de leur analyseur) ou la vitesse de traitement. Très peu s'intéressent aux gains, et aux erreurs supplémentaires que de telles règles peuvent engendrer. Dans cette

section, nous proposons une *évaluation de progression*, qui permet d'appréhender l'impact de chacune des contraintes appliquées sur les règles.

4.1 Les questions d'évaluation

L'objet de notre évaluation est double. Premièrement, nous voulons évaluer la performance de nos règles avec contraintes, c'est-à-dire le pourcentage d'analyses correctes après l'ajout de chaque contrainte. Idéalement, l'ajout de chaque contrainte devrait augmenter la performance de la règle. Le but ultime de chaque règle est de tendre vers une performance maximale, car une règle qui traite les mots inconnus ne devrait pas fournir d'analyse incorrecte (et générer ainsi plus de bruit que le silence qu'elle réduit). Deuxièmement, nous voulons mesurer plus finement les gains de chaque contrainte (les vrais positifs) par rapport aux nombres de « pertes » provoquées par celle-ci, c'est-à-dire le nombre de mots construits « corrects » mais exclus à cause de l'application de la contrainte (les « faux négatifs »).

Pour évaluer l'impact de ces contraintes sur la performance globale de la règle, nous utilisons comme score minimal (la *baseline*) la performance de la règle contrainte, telle qu'elle est présentée à la figure 1, section 3. Dans l'évaluation plus précise des gains et des pertes, nous cherchons à nous approcher des 100% d'analyse correcte pour les « vrais positifs ».

Pour mener à bien l'évaluation, il faut également décider quelle est la bonne réponse, la réponse attendue. En morphologie constructionnelle, et peut-être encore d'avantage en néologie, il est parfois très difficile de dire si un mot est construit ou non. Comme le soulignent (Schmid et al., 2004) à propos des analyseurs morphologiques de l'allemand, « there is no general agreement yet about what constitutes the correct analyses ». Pour notre part, nous considérons qu'une analyse est correcte quand la base et le préfixe sont trouvés pour un mot réellement construit. La difficulté de décider de la bonne analyse dépend de plusieurs facteurs, et notamment la connaissance approfondie de la règle de préfixation. Cette tâche est d'autant plus complexe que les descriptions théoriques sur les préfixes de l'italien ne prennent pas forcément en considération tous les cas de figure, et que l'italien semble être une langue très flexible morphologiquement (nous y reviendrons).

Pratiquement, pour évaluer notre analyseur et ses contraintes, nous lui avons soumis une liste de mots inconnus de notre lexique de référence qui commencent par les mêmes séquences de lettres que les préfixes étudiés. Ces mots ont été extraits d'un important corpus journalistique de l'italien (Baroni et al., 2004, - environ 380 millions d'occurrences). Les occurrences analysées ont ensuite été réduites en formes uniques. Finalement, chaque forme a été évaluée manuellement pour distinguer les mots construits de ceux qui ne l'étaient pas.

4.2 Évaluation des règles sans contrainte

L'évaluation de la performance de l'ensemble des règles implémentées jusqu'à présent dans notre analyseur permet d'obtenir un score de référence (*baseline*) pour le reste de l'évaluation. Comme le montre le tableau 1, l'analyse par cet ensemble de règles a une performance tout à fait honorable. Mais, en distinguant les formes préfixées avec trait d'union de celles qui ne le sont pas, nous remarquons que les règles sont beaucoup moins performantes quand il n'y a pas de trait d'union.

	mots concernés	analyses correctes	analyses incorrectes
avec trait d'union	2839	2833 (99,79 %)	6 (0,21 %)
sans trait d'union	10191	8962 (87,94 %)	1229 (12,05 %)
total	13030	11795 (90,52 %)	1235 (9,48 %)

Tableau 1 : Évaluation des règles sans contrainte

Notons également que la performance n'est pas uniforme entre les règles et dépend beaucoup du préfixe concerné. Ainsi, les préfixes courts ont plus tendance à se retrouver dans des séquences de lettres ambiguës. Par exemple, la règle pour le préfixe *pro* a une performance de 42 %, alors que la règle de *iper* a une performance de 98,29 %. Pour la suite de l'expérience, nous avons pris en compte uniquement les règles de préfixation sans trait d'union, étant donné leur faible performance.

4.3 Évaluation de la règle avec contrainte de catégorie

Globalement, la mise en place de la contrainte de catégorie sur l'ensemble des règles permet d'améliorer légèrement la performance de toutes les règles. Ainsi, d'une performance globale de 87,94 % d'analyses correctes, nous passons à 89,00 %, même si cette variation est différente dans chaque règle. De plus, le nombre de vrais positifs est acceptable, comme le montre le tableau 2.

	total	vrais positifs	faux positifs
analysés	9955	8898 (89,38 %)	1057 (10,62 %)
	total	faux négatifs	vrais négatifs
pas analysés	238	64 (26,89 %)	174 (73,11 %)

Tableau 2 : Ensemble des règles d'analyse avec contrainte de catégorie

Il nous faut également noter que le nombre de mots réellement préfixés mais exclus par les règles (les faux négatifs) est relativement important. Mais ce phénomène varie également beaucoup selon les règles. Par exemple, pour la règle de préfixation en *inter*, nous avions analysé 505 mots, dont 398 étaient réellement des mots construits (performance globale = 78,81 %). L'application de la contrainte catégorielle sur les trois principales catégories lexicales permet d'exclure des séquences de lettres qui avaient été analysées avec une base n'appartenant pas à l'une de ces trois catégories et qui étaient en fait des mots erronés. Le nombre de mots mal analysés a alors diminué, permettant d'augmenter la performance globale de la règle. Le pourcentage de vrais positifs s'élève alors à 79,5 %, sans pour autant exclure des mots réellement construits (0% de faux négatifs). Dans ce cas, l'application de la contrainte, même si elle n'augmente pas significativement la performance, n'exclut pas non plus de mots corrects.

En revanche, avec le préfixe *multi,* nous avons contraint la catégorie de la base aux seuls noms et adjectifs, comme nous l'indique la présentation linguistique faite par

(Iacobini, 2004). Or, l'application abrupte de ce précepte linguistique provoque l'exclusion de dix formes qui étaient réellement des mots construits. Notre lexique avait analysé leur base comme étant des verbes, alors qu'il s'agissait de participes passés employés comme adjectifs. Il convient par conséquent d'être particulièrement prudent avec l'application de certaines « normes » linguistiques et leur adéquation avec les descriptions linguistiques disponibles.

Cette question de « prescription » théorique se retrouve dans le cas particulier du préfixe *mini,* pour lequel les études morphologiques nous indiquent qu'il ne s'accole qu'à des noms pour former des noms (`RCM(mini) : X/NOM = MINI[Y/NOM]`). Nous avons donc exclu toutes les analyses proposant une base non-nominale. Or, si la contrainte de catégorie permet d'obtenir 98,7 % de vrais positifs, elle exclut de l'analyse 22 formes qui sont en fait des mots construits (des faux négatifs). Si ce silence est dû en partie à des mots qui n'avaient pas une base reconnue comme nom par notre lexique de référence, alors qu'elle aurait due l'être (comme dans *un miniporno*, où *porno* est uniquement enregistré comme adjectif), une autre partie concerne des constructions qui ne sont pas nominales (*minigeografica, miniatomica*). La question est de savoir ici comment considérer ces formations. S'agit-il de nouveaux emplois du préfixe selon un usage qui n'est pas encore enregistré par les études linguistiques ? Cette question renvoie à un plus large débat du TALN actuel qui « s'articule désormais entre les règles postulées et les régularités observées » (Habert et Zweigenbaum, 2002, p. 99).

Nous venons de montrer que l'ajustement des règles d'analyse peut améliorer leur performance, mais que les préceptes linguistiques ne doivent pas forcément être pris en compte de manière aveugle, et qu'une évaluation, même partielle, doit en tous les cas être effectuée à chaque nouveau réglage.

4.4 Évaluation de la contrainte de l'indice de constructivité

Concernant l'indice de constructivité des adjectifs relationnels, la performance globale de la règle diminue fortement - 79,37 % alors que la performance pour les règles sur bases adjectivales avant l'application de la contrainte était de 91,38 %. En revanche, comme le montre le tableau 3, l'application de cette contrainte sur les règles de préfixation étudiées (*multi, poli, post, trans* et *inter*) permet d'obtenir une proportion importante de vrais positifs.

	total	vrais positifs	faux positifs
analysés	680	652 (95,88 %)	28 (4,12 %)
	total	faux négatifs	vrais négatifs
pas analysés	202	154 (76,23 %)	48 (23,77 %)

Tableau 3 : Règles d'analyse avec l'indice de constructivité

Si le pourcentage de vrais positifs augmente vraiment, le pourcentage de faux négatifs est très important (ce qui explique la mauvaise performance globale de la règle) et pourrait remettre en cause l'application d'une telle contrainte. Toutefois, une analyse minutieuse des faux négatifs, nous permet, empiriquement cette fois, d'individualiser un certain nombre de constructions typiques semblant favoriser certaines préfixations (comme les mots suffixés en *–ista* et en *-ismo* très fréquemment préfixés en *pro*). De ce constat linguistique, nous passons

alors à des constats qui relèvent davantage d'intuitions de régularité, mais qui sont sans doute valables et qui peuvent donc être ajoutées aux contraintes. Evidemment, de telles intuitions devraient être évaluées sur un second corpus.

Concernant les autres indices de constructivité (sur les noms déverbaux ou sur les adjectifs en *-bile* et les noms en *-tà*), le pourcentage de vrais positifs est très élevé, comme le résume le tableau 4.

	règle sans indice	indice de constructivité	règle avec indice
in + adj	93,14 %	*-bile*	100 %
in + nom	86,70 %	*-tà*	100 %
co + nom	69,48 %	*-(t)ore* | *-zione* | *-mento*	96 %
retro + nom	90 %	*-zione* | *-mento*	100%
ri + nom	91,21 %	*-zione* | *-mento*	99,65 %

Tableau 4 :Performance des règles avec et sans indice de constructivité

Cependant, le nombre de mots exclus par cette contrainte est, on s'en doute, extrêmement important. En effet, la contrainte exclut à chaque fois plus de la moitié des mots commençant par la séquence de lettres concernée, et parmi eux, entre 50 % et 80 % sont réellement construits, ce qui peut évidemment remettre en cause la contrainte de l'indice de constructivité. Il faut néanmoins souligner la performance quasi-optimale des règles ainsi contraintes et donc l'extrême fiabilité de l'analyse produite pour les vrais positifs.

5 Discussion et conclusion

Nous avons montré que de nombreux moyens linguistiques peuvent être mis en place pour améliorer la performance des règles d'analyse des mots construits. Evidemment, des études plus approfondies, tant linguistiques qu'empiriques, permettraient de découvrir d'autres contraintes plus fines pour régler nos règles. Nous avons également montré que pour certaines règles, les indices de constructivité permettaient d'atteindre une performance maximale.

Si toutes ces contraintes augmentent la performance de l'analyse, nous avons également vu qu'il fallait regarder de plus près toutes les conséquences de l'application de ces contraintes. Parfois, nous avons observé que les études linguistiques ne sont pas toujours des sources pertinentes pour la mise en place de ces contraintes, surtout face à la créativité langagière. Nous avons également montré que si certaines contraintes permettent une amélioration considérable de la performance de la règle, elles excluent aussi beaucoup de bonnes analyses. Il est alors important de resituer les objectifs du système pour décider si la perte est plus dommageable que les gains acquis. Dans cette étude, l'application des contraintes permet d'obtenir une performance maximale (100 % de vrais positifs corrects), ce qui, dans notre projet de traduction automatique, est une condition nécessaire à la poursuite du traitement du mot inconnu. Et les pertes importantes (les faux négatifs) peuvent être ici considérées comme un *statu quo* par rapport à leur condition initiale de mots inconnus.

Références

ARONOFF M. (1976) *Word Formation in Generative Grammar*. Cambridge, The MIT press

BARONI M., BERNARDINI S., COMASTRI F., PICCIONI L, VOLPI A., ASTON G., MAZZOLENI M, (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. Acte de *LREC 2004*, 1771-1774.

BYRD R. J. (1983). Word Formation in Natural Language Processing Systems *IJCAI,* 704-706.

BYRD, R. J., KLAVANS J. L., ARONOFF M., ANSHEN F., (1989). Computer methods for morphological analysis. Actes de *24th ACL*, 120-127.

CARTONI B. (2005). Traduction de règles de construction des mots pour résoudre les problèmes d'incomplétude lexicale en traduction automatique Étude de cas. Actes de *RECITAL 2005*, 565-574.

CARTONI B. (2006). Dealing with unknown words by simple decomposition: feasibility studies with Italian prefixes. Actesde *LREC 2006*, 1674-1677.

GRABAR N., TRIBOUT D, DAL G., FRADIN F., HATHOUT N, LIGNON S., NAMER F., PLANCQ C., YVON F., ZWEIGENBAUM P. (2006). Productivité quantitative des suffixations par -ité et -Able dans un corpus journalistique moderne. Actes de *TALN 2006*, 167-175.

GUILBAUD J.-P., BOITET C. (1997). Comment rendre une morphologie robuste du français encore plus robuste en traitant finement les mots inconnus avec les données disponibles. Actes de *TALN'97*,

HABERT B., ZWEIGENBAUM P. (2002) Régler les règles. *TAL* 43(3) 83-105.

IACOBINI C. (2004). I prefissi. *in La formazione delle parole in italiano*. Grossmann M, Rainer F, (éds). Tübingen, Niemeyer: 99-163.

Krott A., Schreuder R., Baayen R. H. (1999) Complex Words in Complex Words *Linguistics* 37(5), 905-926

NAMER F. (2003). Productivité morphologique, représentativité et complexité de la base: le système moQuête. *Langue française* 140, 79-101

PETITPIERRE D., RUSSEL G, (1995). Mmorph, The Multext Morphology. Genève, Issco (Technical Report).

SCHMID H., FITSCHEN A, HEID U. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. Actes de *LREC 2004* 1263-1266

WANDRUSZKA U. (2004). Derivazione aggettivale *in La Formazione delle Parole in Italiano* Grossman M, Rainer F (éds) Tübingen, Niemeyer.

WOODS W. A. (2000). Aggressive morphology for robust lexical coverage. Actes de *Applied natural language processing*