Développement de ressources pour le persan: PerLex2, nouveau lexique morphologique et MElt_fa, étiqueteur morphosyntaxique

Benoît Sagot, Géraldine Walther, Pegah Faghiri, Pollet Samvelian

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

Développement de ressources pour le persan : PerLex 2, nouveau lexique morphologique et MElt$_{fa}$, étiqueteur morphosyntaxique

Benoît Sagot[1] Géraldine Walther[2, 3] Pegah Faghiri[3] Pollet Samvelian[3]
(1) Alpage, INRIA & Univ. Paris 7, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
(2) LLF, CNRS & Univ. Paris 7, 5 rue Thomas Mann, 75205 Paris Cedex 13, France
(3) MII, CNRS & Univ. Paris 3, 27 rue Paul Bert, 94204 Ivry-sur-Seine, France
benoit.sagot@inria.fr, geraldine.walther@linguist.jussieu.fr,
pegah.faghiri@etud.sorbonne-nouvelle.fr, pollet.samvelian@univ-paris3.fr

Résumé. Nous présentons une nouvelle version de PerLex, lexique morphologique du persan, une version corrigée et partiellement réannotée du corpus étiqueté BijanKhan (BijanKhan, 2004) et MElt$_{fa}$, un nouvel étiqueteur morphosyntaxique librement disponible pour le persan. Après avoir développé une première version de PerLex (Sagot & Walther, 2010), nous en proposons donc ici une version améliorée. Outre une validation manuelle partielle, PerLex 2 repose désormais sur un inventaire de catégories linguistiquement motivé. Nous avons également développé une nouvelle version du corpus BijanKhan : elle contient des corrections significatives de la tokenisation ainsi qu'un réétiquetage à l'aide des nouvelles catégories. Cette nouvelle version du corpus a enfin été utilisée pour l'entraînement de MElt$_{fa}$, notre étiqueteur morphosyntaxique pour le persan librement disponible, s'appuyant à la fois sur ce nouvel inventaire de catégories, sur PerLex 2 et sur le système d'étiquetage MElt (Denis & Sagot, 2009).

Abstract. We present a new version of PerLex, the morphological lexicon for the Persian language, a corrected and partially re-annotated version of the BijanKhan corpus (BijanKhan, 2004) and MElt$_{fa}$, a new freely available POS-tagger for the Persian language. After PerLex's first version (Sagot & Walther, 2010), we propose an improved version of our morphological lexicon. Apart from a partial manual validation, PerLex 2 now relies on a set of linguistically motivated POS. Based on these POS, we also developped a new version of the BijanKhan corpus with significant corrections of the tokenisation. It has been re-tagged according to the new set of POS. The new version of the BijanKhan corpus has been used to develop MElt$_{fa}$, our new freely-available POS-tagger for the Persian language, based on the new POS set, PerLex 2 and the MElt tagging system (Denis & Sagot, 2009).

Mots-clés : Ressource lexicale, validation, étiqueteur morphosyntaxique, persan, catégories, PerLex, MElt.

Keywords: Lexical resource, validation, tagger, Persian, POS, PerLex, MElt.

1 Introduction

Les ressources lexicales et les outils de pré-traitement automatique des langues comme les étiqueteurs morphosyntaxiques sont des ressources indispendsables pour développer des ressources plus complexes et progresser rapidement dans la description théorique des langues en donnant accès à un nombre plus conséquent de données. Malheureusement, ils ne sont que trop rarement librement disponibles, et ce même pour des langues ayant un grand nombre de locuteurs et donc de bénéficiaires potentiels. Pour le persan, ce n'est qu'en 2010 que les premiers lexiques librement disponibles ont commencé à apparaître. Notre lexique morphologique PerLex en est un des précurseurs.

Nous avions développé une première version de PerLex (Sagot & Walther, 2010) dont nous proposons desormais une deuxième version partiellement validée. PerLex 2 possède un nouvel inventaire de catégories fondé sur des choix linguistiques discutés au sein du projet ANR/DFG franco-allemand PerGram. Le développement de PerLex 2 s'accompagne de celui d'un étiqueteur morphosyntaxique, MElt$_{fa}$, qui s'appuie sur l'étiqueteur MElt (Denis & Sagot, 2009) et sur une nouvelle version du corpus BijanKhan (BijanKhan, 2004), résultat de corrections significatives de la tokenisation et d'un réétiquetage selon les nouvelles catégories.

Dans cet article, nous exposons les différentes facettes de ce triple travail dans leur succession et leur interaction. Après une rapide présentation des spécificités du traitement du persan, nous décrivons les améliorations de PerLex

effectuées depuis sa première version, les catégories linguistiques retenues et l'inventaire des étiquettes morphosyntaxiques utilisées par MElt_{fa}. Nous détaillons ensuite le travail de retokenisation et de réétiquetage effectué sur le corpus BijanKhan et enfin le développement et l'entraînement de notre étiqueteur morphosyntaxique MElt_{fa}.

2 Le traitement automatique du persan

Le persan, langue indo-européenne du groupe iranien occidental, est parlé par environ 130 millions de locuteurs. Il s'écrit de droite à gauche avec une variante de l'alphabet arabe. Langue iranienne, le persan se distingue par un nombre très réduit de verbes simples (classe fermée de 200 unités environ). La majorité des sens habituellement exprimés par des prédicats sont exprimés par des locutions verbales complexes qui constituent un procédé très productif (Samvelian, 2001), et dont la partie verbale subit une flexion parfois simplifiée par rapport à celle de verbes simples, nécessitant ainsi un traitement et une modélisation particulières. La morphologie nominale du persan affiche assez peu de formes fléchies : nombre, Ézafé (marqueur de dépendance) , déterminant indéfini , marqueur enclitique de définitude optionnel, postposition را- *-râ*. Les adjectifs ne varient qu'en degré. Ils peuvent néanmoins être suivi de l'Ézafé lorsqu'ils suivent un nom modifié ou lorsqu'ils prennent eux-mêmes un objet. En fin de groupe nominal ils peuvent adopter la flexion des noms. La morphologie verbale du persan est légèrement plus complexe. On considère habituellement qu'il y a deux radicaux verbaux en persan, l'un pour les formes du présent, l'autre pour les formes du passé. Autour de ces radicaux se placent les préfixes de temps, aspect et mode et les désinences personnelles. Les préfixes TAM peuvent être précédés du marqueur de négation. Enfin, le persan possède un paradigme de pronoms personnels enclitiques qui se combinent autant avec des noms qu'avec des verbes, des prépositions, des adjectifs et certains adverbes (Lazard *et al.*, 2006).

Le *projet Shiraz* constitue le premier projet important de traitement automatique du persan. Essentiellement consacré à la traduction automatique du persan vers l'anglais (Amtrup *et al.*, 2000), il a également permis la construction d'un lexique bilingue d'environ 50 000 entrées. Il a ensuite été adapté aux outils Xerox pour les automates à états finis (Megerdoomian, 2004). D'autres ressources lexicales électroniques du persan ont également vu le jour ces dernières années. On peut citer la version électronique du *Persian Pronunciation Dictionary* (Deyhime, 2000), un dictionnaire de formes fléchies avec leurs transcriptions phonétiques en accès limité. En mai 2010 a été rendue librement disponible la 4ème version de MULTEXT-East (Erjavec, 2010) qui comporte désormais le persan (QasemiZadeh & Rahimi, 2006). À la même période, a été développé le lexique PerLex dans sa première version (Sagot & Walther, 2010). Outre ces ressources, d'autres outils d'analyse morphologique ou de lemmatisation ont été développés, mais sans conduire à la construction d'un lexique à large couverture (cf. les travaux de Dehdari & Lonsdale (2008), et notamment leur lemmatiseur PerStem, librement disponible). Récemment, ont été développés divers outils et ressources TAL pour le persan, comme des étiqueteurs morphosyntaxiques (QasemiZadeh & Rahimi, 2006; Tasharofi *et al.*, 2007; Shamsfard & Fadaee, 2008), des analyseurs syntaxiques (Hafezi, 2004; Dehdari & Lonsdale, 2008) et des systèmes de traduction automatique (Saedi *et al.*, 2009).

3 PerLex

Dès la première version de PerLex, nous avions développé une description formelle de la morphologie persane dans le formalisme Alexina (Sagot, 2010). Alexina utilise une représentation à deux niveaux qui sépare la description du lexique de son utilisation : un lexique intensionnel (de lemmes) et un lexique extensionnel (de formes), produit automatiquement par *compilation* du lexique intensionnel. Comme détaillé dans (Sagot & Walther, 2010), les entrées lexicales de PerLex 1 proviennent de diverses sources, et notamment du corpus BijanKhan. À l'été 2010, PerLex contenait 35 914 entrées intensionnelles produisant 524 700 entrées extensionnelles pour 494 488 formes distinctes. Des données complémentaires sur PerLex 1 sont indiquées à la table 1 plus bas. Ces données sont mises en regard des données pour la nouvelle version de PerLex, dont nous allons décrire la construction et notamment les étapes de validation et de conversion vers un nouveau jeu de catégories.

La validation du lexique PerLex a été réalisée par deux moyens complémentaires : la comparaison et si possible la fusion avec deux autres ressources, puis une validation manuelle partielle. Pour optimiser le coût de l'étape manuelle, nous avons défini des heuristiques permettant de présenter aux validateurs les entrées lexicales dont la validation semble prioritaire. Ces heuristiques s'appuient notamment sur de nouveaux poids associés aux entrées

lexicales pendant l'étape automatique : pour toutes les catégories autres que les noms et les adjectifs, c'est-à-dire pour les unités lexicales dont nous sommes raisonnablement sûrs de la classe flexionnelle (ou de l'absence de flexion), le simple fait que la forme canonique ait été trouvée dans un des lexiques employés pour la pré-validation a été jugé suffisant pour que l'unité lexicale ne soit pas présentée à la validation manuelle.

La pré-validation a été opérée à l'aide de deux lexiques du persan existants. (1) Le Persian Pronunciation Dictionary (Deyhime, 2000) est un dictionnaire de 23 168 formes fléchies distinctes associées à une ou plusieurs phonétisations possibles (pour un total de 34 967 entrées). Outre qu'il n'est pas librement redistribuable, les entrées lexicales n'y comportent ni catégorie ni lemme. Ce lexique ne nous a donc pas fourni de nouvelles entrées lexicales. Nous l'avons toutefois utilisé pour attribuer un poids supplémentaire à chaque entrée lexicale de PerLex : le poids ajouté au poids précédent d'une entrée lexicale donnée (en général, la valeur par défaut, c'est-à-dire 100) représente le pourcentage de formes fléchies associées par PerLex à cette entrée lexicale qui ont une phonétisation dans le Persian Pronunciation Dictionary. (2) Le lexique persan distribué dans la version 4 de MULTEXT-East (MTE4-fa) (QasemiZadeh & Rahimi, 2006; Erjavec, 2010), ressource libre, nous a permis d'aller plus loin. Il s'agit d'un lexique morphologique dont les 13 006 entrées comportent une forme fléchie, son lemme, une phonétisation de la forme fléchie et du lemme, et une étiquette positionnelle respectant les conventions MULTEXT habituelles, et qui inclut donc naturellement une catégorie. Après avoir défini un tableau de correspondance entre l'inventaire des catégories utilisées par MTE4-fa et celui du corpus BijanKhan et donc de PerLex 1, nous avons converti MTE4-fa dans le formalisme Alexina afin de mettre en œuvre les outils de fusion de lexiques Alexina dont nous disposions. La fusion a permis de rajouter un poids de 100 à toutes les entrées de PerLex 1 ayant fusionné avec une entrée de MTE4-fa, et aussi d'ajouter de nouvelles entrées à PerLex depuis les autres entrées de MTE4-fa. Ces entrées ont toutefois nécessité un travail manuel ultérieur pour leur assigner une classe flexionnelle.

À ce jour, deux campagnes de validation manuelle ont eu lieu. Lors de la première, toutes les entrées lexicales qui n'étaient pas écartées par l'heuristique ci-dessus étaient accessibles dans l'interface de validation. Lors de la seconde campagne de validation, certaines entrées pourtant déjà validées ont été présentées à nouveau, notamment lorsque leur classe flexionnelle avait changé entre temps (naturellement, uniquement des entrées dont la validation n'avait pas indiqué qu'elles étaient entièrement correctes). L'interface de validation est une interface en ligne. Elle utilise une base de données qui stocke à la fois la version de PerLex en cours de validation et l'ensemble des « tickets de validation » déjà enregistrés. Elle se présente sous la forme de tableaux affichant un ensemble de 30 entrées lexicales non encore validées et appartenant à une même catégorie préalablement choisie par le validateur. Chaque ligne correspond à une entrée lexicale, représentée par sa forme canonique et par un ensemble minimum, souvent vide, de formes fléchies permettant d'être certain que la classe flexionnelle associée à l'entrée lexicale est correcte : le validateur n'a pas besoin de connaître le nom ni le contenu des classes flexionnelles utilisées pour valider l'entrée lexicale. Par ailleurs, chaque colonne correspond à un statut attribuable par le validateur à l'unité lexicale, lui permettant d'indiquer la correction ou le type d'erreur concernant une entrée lexicale données. Lors de la deuxième campagne de validation, suite à la prise en compte des résultats de la première, 1097 tickets de validation ont été créés (818 entrées lexicales correctes, 17 catégorie correcte mais une flexion incorrecte, 26 avaient une catégorie incorrecte, 129 totalement erronnées [1] et 11 entrées familières).

Concernant les catégories, la grammaire de référence (Lazard *et al.*, 2006) qui nous avait guidé dans nos premiers travaux de développement de PerLex se limitait aux catégories de substantifs, adjectifs, adverbes, noms de nombre, pronoms, verbes, interjections et particules . Cet inventaire n'était cependant pas assez précis pour le développement d'outils de traitement automatique comme un étiqueteur morphosyntaxique. Nous ne l'avions donc pas suivi. En 2010, notre lexique comportait les étiquettes reprises directement du corpus BijanKhan (BijanKhan, 2004; Amiri *et al.*, 2007) [2], dont certaines nous semblaient néanmoins insuffisantes en termes de pertinence théorique. Afin d'améliorer le lexique et permettre la construction d'un étiqueteur morphosyntaxique compatible autant avec de futures tâches traitement automatique du persan que d'analyse linguistique, nous avons, en un premier temps fixé un inventaire des catégories du persan. Ces catégories sont le fruit d'une réflexion théorique préalable au sein du projet PerGram. Les catégories retenues sont les suivantes : noms, noms propres, adjectifs, adverbes ; verbes, prépositions, conjonctions, classifieurs, pronoms, déterminants, interjections. PerLex 2 comporte désormais un nouveau jeu de catégories en accord avec ces choix théoriques. Nous avons effectué une conversion automatique des anciennes catégories vers notre nouvel inventaire de catégories. Pour les noms (`N`), verbes (`V`), noms propres (`PN`), pronoms (`PRO`), interjections (`INT`) ponctuations (`DELM`) la conversion était directe. Pour les autres catégories, des critères précis ont dû être appliqués manuellement pour redistribuer les mots qui s'y trouvaient vers l'une ou

1. Dont des membres de l'ancienne catégorie `MORP` dans le corpus BijanKhan, voir section 3.
2. `ADJ, ADV, AR, CON, DELM, DET, IF, INT, MORP, MQUA, MS, PN, OH, OHH, P, PP, PRO, PS, QUA, SPEC, N, V`.

l'autre des catégories du nouvel inventaire. Ainsi, les éléments des classes QUA et MQUA sont désormais étiquettés DET (déterminant), ADV ou PRO. La classe MORP qui comprenait des éléments de morphologie constructionnelle a disparue. Dans le corpus, ses éléments ont été recollés à leurs bases au cours des opérations de retokenisation (section 4.1). Une nouvelle catégorie de classifieurs (CLASS) a été ajoutée comportant une partie des éléments précédemment étiquettés SPEC les autres étant désormais dans DET et ADV.

La table 1 oppose PerLex 1 et 2, globalement et pour quelques catégories importantes. Quantativement, différence entre les deux versions du lexique n'est pas très visible, la suppression d'entrées étant allée à l'opposée des ajouts d'entrées manquantes. De plus, le travail de conversion vers le nouveau jeu de catégories, et les autres améliorations (cf. la description morphologique) ne sont pas de nature à se refléter quantitativement.

Partie du discours	entrées intensionnelles		lemmes distincts		entrées extensionnelles	
	PerLex 1	PerLex 2	PerLex 1	PerLex 2	PerLex 1	PerLex 2
verbes	171	176	139	140	19 776	20 373
noms communs	9 553	9 546	9 106	9 073	177 988	165 345
noms propres	10 996	10 965	10 938	10 954	33 076	31 777
adjectifs	11 872	12 322	11 835	12 284	290 537	302 574
autres	3 322	3 706	3 120	3 622	3 323	
total	*35 914*	*36 397*	*33 454*	*35 924*	*524 700*	*525 074*

TABLE 1 – Données quantitatives sur PerLex

4 MElt$_{fa}$: un analyseur morphosyntaxique du persan

PerLex 2 nous a permis de construire un analyseur morphosyntaxique du persan, en bénéficiant de la disponibilité du corpus BijanKhan. Après la définition du jeu d'étiquettes morphosyntaxiques, et avant d'entraîner le système MElt (Denis & Sagot, 2009), nous avons toutefois dû appliquer au corpus BijanKhan des procédures automatiques de correction de corpus (sa qualité en termes de tokenisation et d'étiquetage est insuffisante) et de conversion vers notre jeu d'étiquettes. À partir de l'inventaire de catégories décrit en section 2, nous avons construit un jeu de 79 étiquettes morphosyntaxiques pour notre étiqueteur MElt$_{fa}$. Pour les 12 catégories de notre inventaire, il contient : 37 étiquettes verbales ; 9 pronominales ; 8 nominales ; 5 pour les prépositions ; 3 pour les adjectifs, conjonctions, déterminants et interjections ; 2 pour les adverbes et les classifieurs. Les noms propres ont une étiquette unique. Nous avons également ajouté une étiquette pour les expressions arabes complètes empruntées et citées telles quelles dans les textes persans, une pour les nombres et l'étiquette indispensable pour les ponctuations.

4.1 Amélioration du corpus BijanKhan

Le corpus BijanKhan est un corpus librement disponible de 2 597 937 tokens, résultat d'une tokenisation et d'un étiquetage automatiques de textes journalistiques et généraux. Il n'est segmenté ni en phrases ni en articles. Nous sommes partis de la version translittérée du corpus développée dans (Sagot & Walther, 2010), puis nous l'avons segmenté en 88 885 phrases de façon simple en découpant sur les ponctuations fortes standard. Nous avons alors appliqué un certain nombre de règles systématiques de correction, afin de régler un inventaire assez large d'erreurs et d'incohérences trouvées dans le corpus, dont voici un extrait, par ordre d'application :

– décollage des préverbes (maintenant étiquetés P) des verbes auxquels ils sont parfois collés par erreur, et donc intégrés dans le token verbal, produisant des formes verbales à juste titre inconnues de PerLex ;
– normalisation de la typographie des préfixes flexionnels verbaux : p.ex., le préfixe ب- *be-* doit être lié au caractère suivant, alors que les préfixes می-(ن) *(n)mi-* doivent être collés mais non liés (ni séparé) ;
– restauration de prépositions composées tokenisées par erreur en un « nom » et une préposition (پس + از *pas + az*) ;

Une fois ces corrections effectuées, il nous a encore fallu convertir le corpus BijanKhan de son propre jeu d'étiquettes vers celui décrit à la section 4. Toutefois, étant donné la qualité imparfaite de l'étiquetage du corpus, nous avons fait le choix de nous appuyer autant que possible sur PerLex, ou, parfois, sur certaines généralisations morphologiques (infra), pour « valider » l'étiquette BijanKhan en même temps que de proposer une ou plusieurs étiquettes de notre propre jeu. Il a parfois été impossible de convertir l'étiquette de certains tokens, soit parce que ni PerLex ni les quelques règles morphologiques ne nous la confirment (aucune étiquette n'est proposée), soit au

contraire parce que plusieurs étiquettes sont compatibles. Une fois l'ensemble de des règles de conversions appliquées (nous n'en donnons pas le détail faute de place), certains tokens ont reçu une étiquette venant de notre propre jeu d'étiquettes, d'autres ont conservé celle du BijanKhan. La conversion a pu fonctionner dans 92,4% des cas, produisant ainsi une annotation complète dans notre jeu d'étiquettes pour 18 763 phrases (21% de l'ensemble).

Nous avons alors partagé le corpus en trois parties : 1) les 100 dernières phrases du corpus (dont 32 intégralement converties) ont été séparées des autres en vue de la construction d'un corpus d'évaluation (4.2) ; 2) Les phrases intégralement converties restantes, soit 18 731 phrases (pour 302 690 tokens), ont été rassemblées en un corpus d'entraînement pour MElt ; 3) les autres phrases seront utilisées ultérieurement, notamment pour comparer les annotations créées avec celles que produira MElt$_{fa}$, et pour des tâches d'acquisition automatique d'unités lexicales.

4.2 Entraînement et évaluation de l'étiqueteur morphosyntaxique MElt$_{fa}$

La conversion de PerLex 2 en un lexique simple associant à chaque forme une ou plusieurs étiquettes du jeu d'étiquettes défini en 4 n'a pas posé de problème, PerLex fournissant des informations morphologiques riches, et le jeu d'étiquettes ayant été conçu pour permettre cette conversion. Nous avons donc entraîné le système d'analyse morphosyntaxique MElt (Denis & Sagot, 2009), en lui donnant en entrée le lexique extrait de PerLex 2 et notre corpus d'entraînement de 18 731 phrases. Le résultat est un étiqueteur morphosyntaxique du persan, MElt$_{fa}$.

Pour évaluer cet étiqueteur, nous avons validé et complété manuellement l'étiquetage des 100 dernières phrases du corpus BijanKhan, après qu'elles ont été corrigées et pré-annotées comme le reste du corpus (section précédente). Ce corpus de référence est constitué de 1 707 tokens. La conversion vers notre jeu d'étiquettes a pu s'appliquer pour 1 568 (91,6%) d'entre eux. Nous avons comparé les résultats de MElt$_{fa}$ à ce corpus de référence. et comparés au résultat brut de la correction et pré-annotation, sur les 1 568 tokens dont l'étiquette a été effectivement convertie. Sur l'ensemble de la référence, nous obtenons une précision de 90,3% avec notre jeu de 79 étiquettes, et 93,3% sur les seules catégories (y compris `AR` et `NUM`). Sur les tokens dont l'étiquette a pu être convertie, nous montons à 93,9% et 95,3% respectivement. Ce score constitue certainement une borne inférieure de la précision que nous obtiendrions si tous les tokens étaient convertis : en effet, les tokens non convertis ne l'ont pas été non plus dans les données d'entraînement, et MElt$_{fa}$ n'a donc pu apprendre à leur sujet des informations contextuelles spécifiques, d'où un taux d'erreur supérieur. De plus, ces erreurs sont susceptibles de se répercuter au voisinage de ces tokens.

Nous avons cherché à comparer la qualité des annotations produites par MElt$_{fa}$ à celles résultant de la correction et conversion du corpus BijanKhan — déjà améliorées par rapport au corpus d'origine. Nous avons calculé la précision de ce corpus, restreint aux 1 568 tokens effectivement convertis, par rapport à la référence. Le résultat en précision est exactement identique à celui de MElt$_{fa}$, bien que les erreurs ne concernent les mêmes tokens que dans 48% des cas. Autrement dit, MElt$_{fa}$ a réussi à produire, sur les 91,6% de tokens correctement convertis, un résultat aussi bon que le corpus sur lequel il s'est entraîné, lui même meilleur que le corpus BijanKhan. Nous pensons que ce résultat est lié à la fois à l'utilisation de PerLex, qui l'aide à ignorer certaines données aberrentes du corpus d'apprentissage, et au fait que le modèle produit par MElt$_{fa}$ en lisse nombre d'erreurs (avec un effet de type co-training). Cette dernière hypothèse est confirmée par le fait que sur ces 1 568 tokens, MElt$_{fa}$ produit des résultats légèrement plus proches de la référence (93,9% de précision) que du corpus converti (93,4%).

5 Conclusion

Nous avons développé une nouvelle version de PerLex (Sagot & Walther, 2010) corrigée et modifiée qui, grâce à l'intégration de réflexions théoriques, pourra notamment être plus adaptée aux besoins de ressources pour des travaux d'analyse linguistique. Cette nouvelle version de PerLex repose sur un nouvel inventaire de catégories, a été partiellement validé (semi-automatiquement et manuellement) et corrigé. PerLex est librement disponible sous `http://alexina.gforge.inria.fr`. Nous avons également développé une version corrigée partielle du corpus BijanKhan qui comporte notamment une retokenisation complète. Enfin, nous disposons d'un étiqueteur morphosyntaxique du persan MElt$_{fa}$, librement disponible à l'adresse `http://lingwb.gforge.inria.fr`. Cet étiqueteur complète notre ensemble d'outils de TAL du persan qui comportait déjà la chaîne de pré-traitement SxPipe$_{fa}$ (Sagot & Walther, 2010) librement disponible à la même adresse. Entraîné sur des données bruitées, il est encore améliorable, bien qu'un score de 90,3% sur un jeu de 79 étiquettes soit tout à fait honorable. MElt$_{fa}$ pourra d'ores et déjà être utile pour enrichir PerLex et rechercher des motifs dans des corpus à des fins linguistiques.

Nous comptons, à l'avenir, terminer le volet syntaxique de PerLex, y compris l'intégration des prédicats complexes et de leur structure argumentale. Ce lexique doit être intégré à la grammaire HPSG (Pollard & Sag, 1994) développée au sein de PerGram. Cette grammaire HPSG a vocation a être implémenté dans le système TRALE (Penn, 2004), (Müller & Ghayoomi, 2010). Des travaux d'intégration du formalisme Alexina dans TRALE sont en cours. Une fois l'intégration de PerLex dans TRALE terminée, nous disposerons d'un parser HPSG pour le persan qui pourra notamment servir au développement d'un corpus arboré du persan.

Références

AMIRI H., HOJJAT H. & OROUMCHIAN F. (2007). بررسی پیکره ای مناسب برای برچسب زنی کلمات در زبان فارسی (Investigation on a feasible corpus for Persian POS tagging). In *12th Int. CSI computer conference*, Téhéran, Iran.

AMTRUP J. W., RAD H. M., MEGERDOOMIAN K. & ZAJAC R. (2000). *Persian-English Machine Translation : An Overview of the Shiraz Project.* Memoranda in Computer and Cognitive Science MCCS-00-319, NMSU, CRL.

BIJANKHAN M. (2004). The role of the corpus in writing a grammar : An introduction to a software. *Iranian Journal of Linguistics*, **19**(2).

DEHDARI J. & LONSDALE D. (2008). A Link Grammar Parser for Persian. In S. KARIMI, V. SAMIIAN & D. STILO, Eds., *Aspects of Iranian Linguistics*, volume 1. Cambridge Scholars Press.

DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proc. of PACLIC 2009*, Hong-Kong, China.

DEYHIME G. (2000). *Farhang-i Avayi-i Farsi (Persian Pronunciation Dictionary).* Téhéran, Iran : Farhang Moaser Publishers.

ERJAVEC T. (2010). Multext-east version 4 : Multilingual morphosyntactic specifications, lexicons and corpora. In N. C. C. CHAIR), K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Eds., *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, La Valette, Malte : European Language Resources Association (ELRA).

HAFEZI M. M. (2004). A syntactic parser of Persian sentences. In *Proc. of the 1st Workshop of the Persian Language and Computer*, Téhéran, Iran.

LAZARD G., RICHARD Y., HECHMATI R. & SAMVELIAN P. (2006). *Grammaire du persan contemporain.* Téhéran, Iran : Institut Français de Recherche en Iran & Farhang Moaser Edition.

MEGERDOOMIAN K. (2004). Finite-state morphological analysis of Persian. In *Proc. of the CoLing Workshop on Computational Approaches to Arabic Script-based Languages*, Genève, Suisse.

MÜLLER S. & GHAYOOMI M. (2010). PerGram : A TRALE Implementation of an HPSG Fragment of Persian. In *Proceedings of the International Multiconference on Computer Science and Information Technology*.

PENN G. (2004). Balancing clarity and efficiency in typed feature logic through delaying. In *Proc. of ACL 2004*, p. 239–246.

POLLARD C. & SAG I. (1994). *Head-driven Phrase Structure Grammar.* Stanford, USA : CSLI Publications.

QASEMIZADEH B. & RAHIMI S. (2006). Persian in MULTEXT-East Framework. In *FinTAL*, p. 541--551.

SAEDI C., MOTAZADI Y. & SHAMSFARD M. (2009). Automatic Translation between English and Persian Texts. In *Proc. of the Third Workshop on Computational Approaches to Arabic Script-based Languages*, Ottawa, Canada.

SAGOT B. (2010). The Le*fff*, a freely available, accurate and large-coverage lexicon for French. In *Proc. of the 7th Language Resource and Evaluation Conference*, La Valette, Malte.

SAGOT B. & WALTHER G. (2010). Développement de ressources pour le persan : lexique morphologique et chaîne de traitements de surface. In *Actes de TALN 2010*, Montréal, Canada.

SAMVELIAN P. (2001). Le statut syntaxique des objets nus en persan. *Bulletin de la Société de Linguistique de Paris*, **XCVI**(1), 349--388.

SHAMSFARD M. & FADAEE H. (2008). A hybrid morphology-based pos tagger for Persian. In N. CALZOLARI, Ed., *Proc. of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Maroc.

TASHAROFI S., RAJA F., OROUMCHIAN F. & RAHGOZAR M. (2007). Evaluation of Statistical Part of Speech Tagging of Persian Text. In *International Symposium on Signal Processing and its Applications*, Sharjah, E.A.U.