Développement de ressources pour le persan : le nouveau lexique morphologique PerLex 2 et l'étiqueteur morphosyntaxique MElt-fa

Benoît Sagot, Géraldine Walther, Pegah Faghiri, Pollet Samvelian

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

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

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 & Université 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. Dans cet article 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 : cette nouvelle version 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. In this paper, 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. This new version contains 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 morpho-syntaxique, 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 relativement simples à développer qui, lorsqu'elles sont à la fois de qualité et librement disponibles, permettent néanmoins développer des outils plus complexes comme des analyseurs syntaxiques, des outils de traduction automatique, de fouille de textes et d'extraction d'information pour lesquels elles sont indispensables. Elles permettent également de progresser rapidement dans la description théorique des langues qu'elles traitent en donnant accès à un nombre plus conséquent de données aux chercheurs qui les étudient.

Malheureusement, ces ressources ne sont que trop rarement librement disponibles, et ce même pour des langues importantes comportant un grand nombre de locuteurs et donc de bénéficiaires potentiels. Pour le persan, ce n'est qu'au cours de cette dernière année que les premiers lexiques librement disponibles ont commencé à apparaître. PerLex, notre lexique morphologique du persan, en est un des précurseurs.

Nous avions développé une première version de PerLex (Sagot & Walther, 2010) ; nous en proposons desormais une deuxième version partiellement validée. PerLex 2 possède également un nouvel inventaire de catégories fondé sur des choix linguistiques préalables et motivés. Ces choix ont été faits au sein du projet ANR/DFG franco-allemand PerGram [1]. Le développement de PerLex 2 s'accompagne du développement d'un étiqueteur morpho-syntaxique, MElt_{fa}, qui s'appuie sur lesdits choix linguistiques et le fonctionnement de l'étiqueteur MElt (Denis & Sagot, 2009). Nous avons également développé une nouvelle version du corpus BijanKhan (BijanKhan, 2004) : cette nouvelle version contient des corrections significatives de la tokenisation ainsi qu'un réétiquetage à l'aide des nouvelles catégories. Cette nouvelle version du corpus a ensuite été utilisée pour l'entraînement de MElt_{fa}.

Dans cet article nous exposons les différentes facettes de ce triple travail dans leur succession et leur interaction. Après une rapide présentation des spécificités liées au traitement du persan, nous décrivons les améliorations du lexique PerLex effectuées depuis sa première version, les catégories linguistiques retenues pour le développement de nos outils et ressources et l'inventaire des étiquettes morpho-syntaxiques utilisées par MElt_{fa}. Nous détaillons ensuite le travail de retokenisation et de réétiquetage effectué sur le corpus BijanKhan et en fin le développement et l'entraînement de notre étiqueteur morphosyntaxique MElt_{fa}.

2 Le traitement automatique du persan : état des lieux

Le premier projet de traitement automatique important pour le persan a été le *projet Shiraz*. Consacré à la traduction automatique du persan vers l'anglais (Amtrup *et al.*, 2000), l'un de ses résultats a notamment été la mise en place d'un lexique bilingue d'environ 50 000 entrées [2]. Il s'appuie sur une description du persan au sein d'un modèle de grammaire d'unification (Megerdoomian, 2000) et 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 quatrième 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 n'ont pas conduit à la construction d'un lexique à large couverture. On peut citer les travaux de Dehdari & Lonsdale (2008), et notamment leur lemmatiseur PerStem, librement disponible [3].

Ces dernières années ont été développés divers outils et ressources TAL pour le persan. Ils s'agit essentiellement d'étiqueteurs morpho-syntaxiques (QasemiZadeh & Rahimi, 2006; Tasharofi *et al.*, 2007; Shamsfard & Fadaee, 2008), d'analyseurs syntaxiques (Hafezi, 2004; Dehdari & Lonsdale, 2008) et de systèmes de traduction automatique (Feili & Ghassem-Sani, 2004; Saedi *et al.*, 2009).

3 Le persan

Le persan, langue indo-européenne de la famille des langues indo-aryennes, appartient plus précisément au groupe des langues iraniennes occidentales qui se caractérise par un ordre des mots relativement libre mais avec une préférence pour le type SOV. Il est parlé par environ 130 millions de locuteurs, répartis surtout en Iran, en Afghanistan, au Tadjikistan et en Ouzbékistan.

1. PerGram a pour but (1) d'établir une modélisation HPSG d'un certain nombre de phénomènes syntaxiques du persan (comme celui des prédicats complexes), (2) d'implémenter une grammaire HPSG couvrant l'essentiel des phénomènes linguistiques du persan et (3) de mettre en place des ressources et outils automatiques. PerLex est prévu pour être associé en tant que lexique à la fois morphologique et syntaxique à la grammaire HPSG de PerGram. PerGram est un projet dirigé conjointement par Pollet Samvelian (Université Paris 3) et Stefan Müller (Freie Universität Berlin).
2. Malheureusement, ce lexique ne semble pas être librement disponible.
3. `http://sourceforge.net/projects/perstem/`

3.1 Écriture et translittération

Le persan s'écrit de droite à gauche au moyen d'une variante de l'alphabet arabe. De même qu'en arabe, les voyelles brèves ne sont pas transcrites et la distinction entre majuscules et minuscules n'existe pas. Par ailleurs, deux caractères consécutifs peuvent être *liés* (c'est-à-dire écrits d'un seul trait, ce qui n'est possible que pour certains caractères), *collés* (juxtaposés sans être liés) ou séparés par un espace. Selon qu'il est isolé ou lié au caractère précédent et/ou suivant, un même caractère peut prendre jusqu'à quatre formes différentes.

Certains outils et ressources développés au cours de notre travail reposent sur une translittération du persan en caractères latins. Nous avons utilisé la translittération bijective développée dans le cadre du projet PerGram [4].

En plus d'un outil de translittération permettant de basculer d'un alphabet (persan ou latin) à l'autre, nous avions aussi au préalable développé deux outils de normalisation typographique (Sagot & Walther, 2010).

3.2 Aperçu de la grammaire du persan

Comme la plupart des langues iraniennes, le persan se distingue par un nombre très réduit de verbes simples (classe fermée de 200 unités environ seulement). Dans les langues iraniennes, la majorité des sens habituellement exprimés par des prédicats sont exprimés par des locutions verbales complexes qui constituent un procédé très productif (Samvelian, 2001). Dans ces locutions verbales complexes, la partie verbale subit une flexion parfois simplifiée par rapport à celle de verbes simples, nécessitant ainsi un traitement et une modélisation particulières (Lazard *et al.*, 2006).

La morphologie nominale du persan n'affiche qu'un nombre restreint de formes fléchies (Lazard *et al.*, 2006) : nombre (singulier non marqué, pluriel en ـها / ـان *-hâ/-ân* et plus rarement *-ât/-un/-in*), Ézafé (marqueur de dépendance) ـی *-e/-ye* (Samvelian, 2007), déterminant indéfini ـی *-î*, marqueur enclitique de définitude optionnel *-h(e)*, postposition ـرا *-râ*. Les adjectifs ne varient qu'en degré (suffixe ـتر *-tar* au comparatif et suffixe ـترین *-tarin* au superlatif). Ils peuvent néanmoins être suivis de l'Ézafé lorsqu'ils suivent un nom modifié ou lorsqu'ils prennent eux-mêmes un objet direct ou indirect. 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 du persan, l'un pour les formes du présent, l'autre pour les formes du passé. Autour de ces radicaux se placent des affixes : des préfixes TAM [5] می- *mi-* et ب- *be-* et les désinences personnelles *-am*, *-i*, *-ad/-e/ø*, *-im*, *-id/-in* et *-and/-an* [6] (ou les formes enclitiques du verbe بدن *budan* 'être'. Les préfixes TAM peuvent être précédés du marqueur de négation ن- / م- *n-/m-*. Les formes du parfait et de l'imparfait sont périphrastiques et construites à partir du participe et des formes enclitiques de l'auxiliaire بدن *budan* 'être'.

Enfin, le persan possède un paradigme de pronoms personnels enclitiques qui peuvent se combiner aussi bien avec des noms qu'avec des verbes, des prépositions, des adjectifs et certains adverbes (Lazard *et al.*, 2006).

3.3 Inventaire des catégories du persan

Afin d'améliorer à la fois le lexique et permettre la construction d'un étiquetteur morpho-syntaxique 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 [7] :

Classes ouvertes : noms, noms propres, adjectifs, adverbes ;

Classes fermées : verbes, prépositions, conjonctions, classifieurs, pronoms, déterminants, interjections.

4. Pour plus de lisibilité, nous employons ici une transcription phonétique plus standard comprenant aussi les voyelles brèves.
5. Temps-Aspect-Mode.
6. ـد ،ـی- ،ـم- / ـی- , -ø, ـد ،ـم- / ـند- ،ـن .
7. Une description plus exhaustive précisant notamment les choix linguistiques est actuellement en préparation (Samvelian et Faghiri, en préparation).

4 PerLex

Dans cette section, nous décrivons le développement de la nouvelle version de notre lexique PerLex. Nous rappelons l'état de PerLex avant les nouveaux travaux et présentons les diverses améliorations, modifications et validations qui ont mené à la version actuelle de PerLex.

4.1 PerLex 1

Dès la première version de PerLex, nous avions développé une description formelle de la morphologie persane dans le formalisme Alexina. Alexina permet de représenter les informations lexicales d'une façon complète, efficace et lisible (Sagot, 2005; Danlos & Sagot, 2008), tout en étant compatible avec la norme ISO pour les lexiques TAL, la norme LMF (Lexical Markup Framework) (Francopoulo *et al.*, 2006). D'autres ressources pour d'autres langues sont également développées dans ce formalisme [8].

Alexina utilise une représentation à deux niveaux qui sépare la description du lexique de son utilisation :
– le lexique intensionnel factorise les informations lexicales en associant à chaque lemme une classe morphologique (définie dans une description morphologique formalisée) et des informations syntaxiques profondes ; il est utilisé pour le développement de la ressource ;
– le lexique extensionnel, produit automatiquement par *compilation* du lexique intensionnel, associe à chaque forme fléchie une structure détaillée qui représente l'ensemble de ses propriétés morphologiques et syntaxiques ; il est directement utilisé par les outils TAL tels que les étiqueteurs ou les parseurs.

Notre description morphologique du Persan à partir des données de (Lazard *et al.*, 2006) dans le formalisme morphologique d'Alexina (Sagot, 2007) modélise l'affixation comme la combinaison d'un préfixe et d'un suffixe avec un radical. Les alternances de radicaux sont pour l'instant modélisées comme des phénomènes de *sandhi* [9] affectant le radical à ses frontières avec les affixes. Dans le formalisme Alexina, ces sandhi peuvent également affecter les affixes directement.

PerLex utilise aujourd'hui entre autres 31 tables verbales et 5 tables nominales.

Le formalisme Alexina couvre à la fois le niveau morphologique et le niveau syntaxique (p.ex. la valence). Une formalisation de la syntaxe du persan est actuellement en cours. Elle s'appuie notamment sur les choix en termes de structure argumentale fait au sein du projet PerGram (Faghiri, en préparation). Les tables de descriptions linguistiques exhaustives des structures argumentales des verbes simples seront ensuite converties dans le format syntaxique d'Alexina. La poursuite de travaux descriptifs sur les prédicats complexes (Samvelian, 2011; Samvelian & Jorgensen, 1993) est également prévue, puis l'intégration de ces unités lexicales particulières dans PerLex. De façon complémentaire et pour viser une couverture optimale de la ressource, des techniques d'extraction automatique de prédicats complexes candidats, puis, à terme, de leur structure argumentale, est en cours au sein du projet PerGram (Gutman, Sagot & Samvelian, en préparation).

Comme détaillé dans (Sagot & Walther, 2010), les entrées lexicales de PerLex 1 ont été extraites à partir de diverses sources, et notamment à partir du corpus BijanKhan. À l'été 2010, PerLex contenait 35 914 entrées intensionnelles (de niveau lemme) produisant 524 700 entrées extensionnelles (de niveau forme) pour 494 488 formes distinctes. Des données complémentaires sur PerLex 1 sont indiquées au tableau 1, à la fin de la section suivante. 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.

8. Le Lefff, un lexique morphologique et syntaxique à large couverture pour le français (Sagot, 2010), ainsi que des ressources pour l'espagnol (le Leffe), le galicien, le polonais (Sagot, 2007), le slovaque (Sagot, 2005), et l'anglais ; il y a aussi deux lexiques pour des langues kurdes, langues iraniennes typologiquement proches du persan : SoraLex, un lexique morphologique pour le kurde sorani (Walther & Sagot, 2010) et KurLex, lexique morphologique du kurde kurmanji (Walther *et al.*, 2010).

9. Transformations sur le radical et/ou l'affixe provoquées par la juxtaposition ce ces derniers. Ainsi en français, lorsque le préfixe *in-* est juxtaposé aux bases dérivationnelles *réversible* ou *légal*, un phénomène de *sandhi* a lieu qui induit la modification de la consonne finale *n* du préfixe, créant ainsi les formes *ir-réversible* et *il-légal* (à la différence de *in-attendu*).

4.2 Améliorations, validation et nouvel inventaire de catégories : PerLex 2

4.2.1 Enrichissement et pondération automatique par comparaison et fusion avec d'autres lexiques

La validation du lexique PerLex a été réalisée par deux moyens complémentaires : la comparaison et si possible la fusion avec deux autres ressources (le lexique MULTEXT-East et le Persian Pronunciation Dictionary), puis une validation manuelle partielle.

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 donc 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 l'entrée lexicale qui ont une phonétisation dans le Persian Pronunciation Dictionary.

Le lexique persan distribué dans la version 4 de MULTEXT-East (désormais MTE4-fa) (QasemiZadeh & Rahimi, 2006; Erjavec, 2010), qui est une ressource libre, nous a permis d'aller plus loin. En effet, 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 utilisé par le corpus BijanKhan et donc par PerLex 1, nous avons converti MTE4-fa dans le formalisme Alexina, pour pouvoir mettre en œuvre les outils de fusion de lexiques Alexina dont nous disposions (Molinero *et al.*, 2009). La fusion a alors permis de rajouter un poids de 100 à toutes les entrées de PerLex 1 ayant fusionné avec une entrée de MTE4-fa, mais également d'ajouter de nouvelles entrées à PerLex 1 à partir des entrées de MTE4-fa n'ayant pas fusionné. Ces entrées ont toutefois nécessité un travail manuel ultérieur, afin de leur assigner une classe flexionnelle.

4.2.2 Validation manuelle

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 les nouveaux poids attribués à l'étape précédente : pour toutes les catégories autres que les noms et les adjectifs, c'est-à-dire pour les catégories dont 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 MTE4-fa ou dans le Persian Pronunciation Dictionary suffit à ce que l'unité lexicale ne soit pas présentée à la validation.

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, après prise en compte des résultats de la première, certaines entrées pourtant déjà validées ont été présentées à nouveau, notamment lorsque leur classe flexionnelle avait changé entre temps (naturellement, ceci n'a donc concerné que 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. Comme le montre la figure 1, l'interface 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, qui est représentée par sa forme canonique et par un ensemble minimum, souvent vide, de formes fléchies : il s'agit du plus petit ensemble de formes fléchies permettant d'être certain que la classe flexionnelle associée à l'entrée lexicale est correcte. Ainsi, le validateur n'a pas besoin de connaître le nom et le contenu des classes flexionnelles utilisées pour valider l'entrée lexicale. Par ailleurs, chaque colonne correspond à un statut attribué par le validateur à l'unité lexicale. Lors de la première campagne, les statuts possibles étaient simplement « correct », « catégorie correcte mais flexion incorrecte », « lemme valide mais catégorie incorrecte » et « lemme invalide ». L'objectif était de rendre la validation aussi rapide que possible.

Toutefois, il s'est avéré que ce choix était trop extrême, et que le statut « catégorie correcte mais flexion incorrecte » gagnerait à être répartie entre plusieurs statuts possibles, notamment pour les noms. En effet, en persan, un nom se

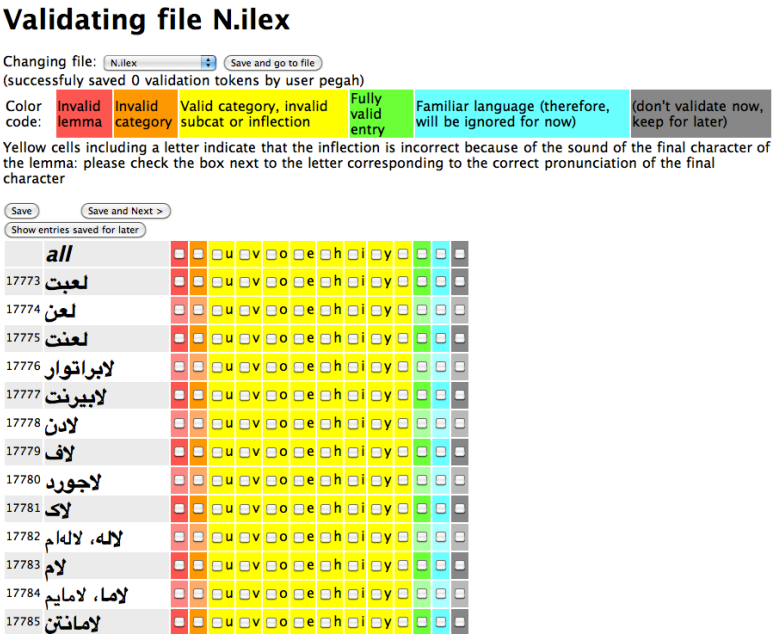

Figure 1 – Interface de validation telle qu'utilisée pour la deuxième campagne

terminant par la lettre و *vâv* se fléchit différemment selon que cette lettre se prononce [u], [v] ou [o]. De même, un nom se terminant par un ه *heh* se fléchit différement selon qu'il se prononce [h] (consonne) ou [e] (voyelle). Il en va de même avec la lettre ی *yeh*, prononcée [i] ou [j]. À chacun de ces sept cas correspond désormais une colonne distincte, qui signifie « catégorie correcte, flexion incorrecte, car la dernière lettre devrait se prononcer ainsi ». La colonne d'origine est préservée, pour indiquer les autres cas de flexion incorrecte.

La première campagne de validation a également permis d'identifier des fonctionnalités utiles, ajoutées avant le démarrage de la seconde campagne. Ainsi, un panier d'entrées lexicales laissées temporairement de côté a été créé, de même qu'un nouveau statut, « entrée lexicale familière » [10]. Sont également prévus le développement d'un outil de recherche ainsi que la possiblité d'afficher des exemples en dessous de chaque entrée lexicale à valider — notamment dans la perspective de la validation des futures entrées pour les prédicats complexes.

Lors de la première campagne de validation, 751 tickets de validation ont été créés, principalement sur des entrées nominales. Au total, 451 entrées lexicales ont été jugées correctes, 250 avec une catégorie correcte mais une flexion incorrecte, aucune avec une catégorie incorrecte, et 50 totalement incorrectes (mais dans un grand nombre de cas c'était la conséquence d'un bogue sur certains caractères particuliers, résolu depuis).

Lors de la deuxième campagne de validation, qui a suivi la prise en compte des résultats de la première (tickets de validation, mais aussi erreurs systématiques identifées), 1097 tickets de validation ont été créés, sur diverses catégories. Au total, 818 entrées lexicales ont été jugées correctes, 17 ont reçu un des 8 statuts indiquant une catégorie correcte mais une flexion incorrecte, 26 avaient une catégorie incorrecte, 129 étaient totalement erronnées (dont de nombreuses formes fléchies de pronoms considérées par erreur comme des entrées indépendantes [11]) et 11 étaient des entrées familières.

4.2.3 Conversion de PerLex vers le nouvel inventaire de catégories

La grammaire de référence (Lazard *et al.*, 2006) qui nous avait guidé dans nos premiers travaux de développement du lexique du persan PerLex se limitait aux catégories suivantes : substantifs, adjectifs, adverbes, noms de nombre, pronoms, verbes, interjections et particules (prépositions, conjonctions, particules adverbiales, interrogatives et exclamatives). Cet inventaire n'était cependant pas assez précis pour le développement d'outils de traitement

10. Le projet PerGram a fait le choix d'inclure dans le lexique élaboré uniquement les entrées de la langue standard (bien que les formes fléchies familières des verbes standard soient produites).
11. Cf. des membres de l'ancienne catégorie MORP dans le corpus BijanKhan, voir section 4.2.3.

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) [12], dont certaines nous semblaient néanmoins insuffisantes en termes de pertinence théorique, voire complètement opaques.

La nouvelle version de PerLex comporte désormais un nouveau jeu de catégories en accord avec les choix théoriques mentionnés en 3.3. 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 l'autre des catégories du nouvel inventaire. Ainsi, les classes `QUA` et `MQUA` comportaient des éléments qui 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 (cf. section 5.2). À l'inverse, une nouvelle catégorie de classifieurs (`CLASS`) a été ajoutée. Elle contient une partie des éléments précédemment étiquettés `SPEC`. Les éléments de l'ancienne catégorie `SPEC` se retrouvent desormais dans les catégories `CLASS`, `DET` et `ADV`.

4.3 PerLex : données quantitatives

Le tableau 1 met en regard quelques chiffres sur PerLex 1 et 2, à la fois au niveau global et pour quelques catégories importantes. On constate que la différence entre les deux versions du lexique n'est pas très visible au niveau des chiffres. D'une part, la suppression d'entrées est allée à l'opposée des ajouts d'entrées manquantes. D'autre part, le travail de conversion vers le nouveau jeu de catégories, et de nombreuses autres améliorations (notamment dans la description morphologique [13]) ne se reflètent pas quantitativement de façon significative.

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

PerLex 2 reste toutefois un lexique d'échelle raisonnable. Calculer sa couverture sur le corpus BijanKhan n'a pas grand sens, dans la mesure où une partie importante des entrées lexicales en ont été extraites, mais également parce que la segmentation de ce corpus en unités lexicales n'est pas parfaite. Le rôle de PerLex 2 pour la construction de notre analyseur morphosyntaxique du persan MElt$_{fa}$ en permet cependant une évaluation indirecte.

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

Cette nouvelle version de PerLex nous a permis de construire un analyseur morphosyntaxique du persan, en bénéficiant de la disponibilité du corpus BijanKhan. Après la définition d'un jeu d'étiquettes morphosyntaxiques, et avant d'entraîner le système MElt (Denis & Sagot, 2009), nous avons dû toutefois appliquer au corpus BijanKhan des procédures automatiques de correction de corpus (sa qualité est loin d'être excellente, tant en termes de tokenisation que d'étiquetage) et de conversion vers notre jeu d'étiquettes.

12. L'inventaire de base est le suivant : `ADJ, ADV, AR, CON, DELM, DET, IF, INT, MORP, MQUA, MS, PN, OH, OHH, P, PP, PRO, PS, QUA, SPEC, N, V`.

13. Par exemple, trois nouvelles tables de flexion ont été rajoutées pour mieux rendre compte de la flexion particulière des verbes support (cf. section 3.2) et des auxiliaires.

5.1 Définition d'un jeu d'étiquettes morphosyntaxiques

À partir de l'inventaire de catégories décrit en section 3.3, nous avons construit un jeu de 79 étiquettes morphosyntaxiques pour notre étiqueteur MElt_{fa}. Ce jeu d'étiquettes est résumé dans le tableau 2. Pour les douze catégories de notre inventaire, il contient : 37 étiquettes verbales ; 9 étiquettes pronominales ; 8 étiquettes nominales ; 5 étiquettes 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.

5.2 Retokenisation et correction 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 [14] ;
– 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) ;
– correction de diverses erreurs typographiques similaires concernant certains suffixes verbaux ;
– correction des erreurs typographiques liées à la marque du pluriel nominal ها- *-hâ* possiblement suivie de marques d'indéfini ou personnelles [15] ;
– correction de diverses autres erreurs typographiques similaires concernant les noms et les adjectifs ;
– tentative de normalisation typographique des sigles ;
– restauration de certaines prépositions composées tokenisées par erreur en un « nom » et une préposition (cf. پس + از *pas + az*) ;
– prise en compte de la disparition de la catégorie `MORP`, présente dans le BijanKhan mais absente de notre inventaire de catégories ; ceci a notamment conduit à regrouper en un seul token et à identifier comme tels un large éventail d'expressions numériques ;
– quelques corrections systématiques d'étiquetage ;
– quelques retokenisations de tokens erronés, notamment certains fusionnant un mot et une ponctuation.

5.3 Conversion du corpus vers le nouveau jeu d'étiquettes

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 5.1. 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 (cf. infra), pour « valider » l'étiquette BijanKhan en même temps que de proposer une ou plusieurs étiquettes de notre propre jeu. Il est arrivé que nous ne puissions 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.

Nous ne décrivons pas ici le processus complet de conversion, mais nous en donnons néanmoins quelques exemples :

– les formes étiquetées par certaines classes telles que `ADV` (adverbe), `AR` (emprunt à l'arabe), `PRO` (pronom) reçoivent l'étiquette proposée par PerLex pour cette forme, à condition que PerLex en propose exactement une (c'est-à-dire que PerLex contient ce mot avec cette catégorie, mais qu'il n'y a aucune ambiguïté sur l'étiquette, par exemple aucune ambiguïté entre `PROpers`, `PROrefl`, `PROrecip`, `PROdem` etc.)

14. Exemples de tels préfixes : بر *br*, در *dr*, باز *baz*, فرا *fra*, پیش *pyš*, *sr*. On notera que certaines formes verbales commencent réellement par *br-*, *dr-* ou *baz-*, ce qui nécessite des exceptions à cette règle.
15. Y compris certaines corrections, comme le changement systématique du suffixe هاای- *-eaay* en های- *-eay*.

cat.	étiquette	description	critères	exemple
DELM	DELM		*signes de ponctuation*	, . ? ... ! ;
N	Nsing	noms au singulier	*noms nus*	افسانه
	Npl	noms au pluriel	*marques flexionnelles (désormais MF) du pluriel*	افسانه-ها
	Nsing_indef	noms au singulier indéfini	*marque -y de l'indéfini*	افسانه-ای
	Npl_indef	noms au pluriel indéfini	*MF du pluriel et de l'indéfini*	افسانه-ها-ی
	Nsing_pers	noms au singulier possessif	*pronoms personnels enclitiques*	افسانه-ام
	Npl_pers	noms au pluriel possessif	*MF du pluriel et enclitiques personnels*	افسانه-ها-یم
	Nvoc	noms au vocatif	*se terminent en -â ou commencent par êy-/yâ-*	پروردگار-ا
	Nunit	signes mathématiques	*liste exhaustive*	kcal
PN	PN	noms propres	*pas de marques de pluriel*	ایرلند
ADJ	ADJ	adjectifs	*peuvent être précédé de kheili 'très'*	بزرگ
	ADJcomp	adjectifs au comparatif	*marque du comparatif -tar*	بزرگ-تر
	ADJsup	adjectifs au superlatif	*marque du superlatif -tarin*	بزرگ-ترین
ADV	ADV	adverbes	*se terminent en -an*	عملا
			liste d'emprunts arabes	یحتمل
			autres : liste exhaustive	امروز
	AR	expression arabes complétes	*emprunts figés*	بسم الله الرحمان الرحیم
DET	DETdem	déterminants démontratifs	*liste exhaustive*	همان
	DETinter	déterminants interrogatifs	*liste exhaustive*	کدام
	DETquant	déterminants quantifieurs	*liste exhaustive*	بعضی
INT	INT	interjections	*liste exhaustive*	شر شر
	INTexcl	interjections exclamatives	*liste exhaustive*	خوشا
	INTs	mots phrase	*liste exhaustive*	صبح بخیر
CLASS	CLASSsing	classifieurs singulier	*peut être situé entre čand et un Nsing*	دانه
	CLASSpl	classifieur pl	*unique*	تا
CONJ	CONJcoord	conjonctions de coordination	*liste exhaustive*	آما
	CONJsubord	conjonctions de subordination	*liste exhaustive*	هرچند که
	CONJcond	conjonctions conditionnelles	*expressions composées avec agar*	اگر چه
PRO	PROpers	pronoms personnels	*liste exhaustive*	او
	PROindef	pronoms indéfinis	*liste exhaustive*	کسی
	PROquant	pronoms quantifieurs	*liste exhaustive*	کلیه
	PROquindef	pronoms qunatifieurs indéfinis	*liste exhaustive*	یکی
	PROinter	pronoms interrogatifs	*liste exhaustive*	کدامیک
	PROdem	pronoms démonstratifs	*liste exhaustive*	همان
	PROrefl	pronoms réfléchis	*liste exhaustive*	خود
	PROrecip	pronoms réciproques	*liste exhaustive*	یکدیگر
	PROnum	pronoms numéraux	*liste exhaustive*	سومی
P	P	prépositions simples	*liste exhaustive*	تا
	Ppn	prépositions complexes P-N	*séquence préposition-nom-ézafé*	بر خلاف
	Pnp	prépositions complexes N-P	*séquence nom-ézafé-préposition*	بنا بر
	Ppp	prépositions complexes P-P	*séquence nom-ézafé-préposition*	به جز
	P+PRO	préposition + PROpers enclitique	*séquence bê/bâ + h(â) + PROpers enclitique*	به-م
	NUM	nombres	*nombres en chiffres ou en lettres*	هفت
V	Vinf	verbes à l'infinitif	*radical 2 + -n*	افشرد-ن
	Vinf_neg	verbes	*m/n-radical-n*	نی-افشرد-ن
	Vinf_apo	verbes à l'infinitif apocopé	*radical 2*	افشرد
	Vpart_pa	verbes au participe passé	*radical 2+ a*	افشرد-ه
	Vpres	verbes à l'imparfait	*mî + radical +*	می-افشار-یم
	Vpastprog	verbes à l'imparfait	*mî-radical 2 + désinences personnelles*	می-افشرد-یم
	Vpastprog_fam	formes familières d'imparfait	*mî-radical 2 + dés. pers. familières*	می-افشرد-ن
	Vpastprog_fam_neg	formes familières d'imparfait négatif	*n-mî-radical 2 + dés. pers. familières*	ن-می-افشرد-ن
	etc.			

Table 2 – Étiquettes morpho-syntaxiques retenues

– pour les formes verbales, que le BijanKhan étiquette au moyen d'un inventaire restreint (`V_PRS` pour les présents, `V_SUB` pour les subjonctifs, `V_PA` pour les passés, etc.), on vérifie si la forme est connue de PerLex en tant que forme verbale avec une étiquette morphologique compatible avec l'étiquette du BijanKhan ; si c'est le cas et que cela conduit au choix d'une seule étiquette, cette étiquette est choisie ;
– si PerLex ne propose qu'une seule étiquette de nom singulier pour une forme étiquetée `N_SING`, on la choisit ; de même pour le pluriel ; si PerLex ne connaît pas comme une forme nominale plurielle une forme étiquetée `N_PLUR` mais que celle-ci se termine en ات- *-at*, on lui attribue l'étiquette `Npl` ;
– le mot را- *-ra* reçoit l'étiquette `RA`, quelle que soit son étiquette dans le BijanKhan.

Une fois l'ensemble de ces règles appliquées, certains tokens ont reçu une étiquette venant de notre propre jeu d'étiquettes, d'autres ont conservé l'étiquette du BijanKhan. La conversion a pu fonctionner dans 92,4% des cas [16], produisant ainsi une annotation complète dans notre jeu d'étiquettes pour 18 763 phrases (soit 21% de l'ensemble).

Nous avons alors partagé le corpus en trois parties :
– 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 (cf. 5.4) ;
– à partir du reste du corpus, les phrases intégralement converties, soit 18 731 phrases (pour 302 690 tokens), ont été rassemblées en un corpus d'entraînement pour MElt ;
– les autres phrases seront utilisées ultérieurement, notamment pour réaliser une comparaison des annotations créées avec celles que produira MElt_{fa}, ainsi que pour des tâches d'acquisition automatique d'unités lexicales.

5.4 Entraînement et évaluation de l'étiqueteur morpho-syntaxique 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 5.1 n'a pas posé de problème, PerLex fournissant d'une part des informations morphologiques riches, et le jeu d'étiquettes ayant d'autre part été conçu pour permettre une telle conversion.

Nous avons donc entraîné le système d'analyse morphosyntaxique MElt (Denis & Sagot, 2009), en lui donnant en entrée le lexique ainsi extrait à partir de PerLex 2 et le corpus d'entraînement de 18 731 phrases mentionné ci-dessus — bien qu'il ne s'agisse pas d'un corpus validé manuellement, et qu'il contienne donc de nombreuses erreurs. 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 (cf. 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. Nous les avons également 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. Sur les tokens dont l'étiquette a pu être convertie, nous montons à 93,9% sur le jeu d'étiquettes et 95,3% sur les 14 catégories [17]. Ce score constitue vraisemblablement 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 — et qui sont donc déjà améliorées par rapport au corpus d'origine. Nous avons donc 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,

16. Nos premiers résultats étaient moins bons, mais l'analyse systématique des cas de non-conversion les plus fréquents nous a permis de corriger et de compléter PerLex — y compris des points précis de la description morphologique formalisée du persan sur laquelle il repose.
17. Y compris `AR` et `NUM`.

MElt$_{fa}$ produit des résultats légèrement plus proches de la référence (93,9% de précision) que du corpus converti (93,4%).

6 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 validée (semi-automatiquement et manuellement) et corrigée. 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 morpho-syntaxique 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 (Meurers *et al.*, 2002; Penn, 2004; Müller, 2007), (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).

Danlos L. & Sagot B. (2008). Constructions pronominales dans dicovalence et le lexique-grammaire –- intégration dans le Le*fff*. In *Proceedings of the 27th Lexicon-Grammar Conference*, L'Aquila, Italie.

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 *Proceedings 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., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, La Valette, Malte : European Language Resources Association (ELRA).

Feili H. & Ghassem-Sani G. (2004). An Application of Lexicalized Grammars in English-Persian Translation. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*, Valence, Spain.

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M. & Soria C. (2006). Lexical Markup Framework (LMF). In *Proceedings of LREC'06*, Gênes, Italie.

Hafezi M. M. (2004). A syntactic parser of Persian sentences. In *Proceedings 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. (2000). Unification-based Persian morphology. In A. Gelbukh, Ed., *Proceedings of CICLing 2000*, Mexico.

Megerdoomian K. (2004). Finite-state morphological analysis of Persian. In *Proceedings of the CoLing Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland.

Meurers W. D., Penn G. & Richter F. (2002). A web-based instructional platform for constraint-based grammar formalisms and parsing. In D. Radev & C. Brew, Eds., *Effective Tools and Methodologies for Teaching NLP and CL*, p. 18–25, New Brunswick, NJ : The Association for Computational Linguistics. Proceedings of the Workshop held at the 40th Annual Meeting of the Association for Computational Linguistics. 7.–12. July 2002. Philadelphia, PA.

Molinero M. A., Sagot B. & Nicolas L. (2009). A morphological and syntactic wide-coverage lexicon for Spanish : The Leffe. In *Proceedings of the 7th conference on Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgaria.

Müller S. (2007). The Grammix CD Rom. a software collection for developing typed feature structure grammars. In T. H. King & E. M. Bender, Eds., *Grammar Engineering across Frameworks 2007*, Studies in Computational Linguistics ONLINE, p. 259–266. Stanford : CSLI Publications.

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 *Proceedings 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 *Proceedings of the Third Workshop on Computational Approaches to Arabic Script-based Languages*, Ottawa, Ontario, Canada.

Sagot B. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 ((c) Springer-Verlag), Proceedings of TSD'05*, p. 156–163, Karlovy Vary, Czech Republic.

Sagot B. (2007). Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In *Proceedings of the 3rd Language & Technology Conference (LTC'05)*, p. 423–427, Poznań, Poland.

Sagot B. (2010). The Lefff, a freely available, accurate and large-coverage lexicon for French. In *Proceedings 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.

Samvelian P. (2007). A phrasal affix analysis of the Persian Ezafe. *Journal of Linguistics*, **43**(3), 605--645.

Samvelian P. (2011). *Les prédicats complexes Nom-Verbe en persan. Étude syntaxique et sémantique*. Hermès-Lavoisier.

Samvelian P. & Jorgensen H. (1993). Support verb constructions in HPSG. Manuscrit non publié, Université Paris VII.

Shamsfard M. & Fadaee H. (2008). A hybrid morphology-based pos tagger for Persian. In N. Calzolari, Ed., *Proceedings 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.

Walther G. & Sagot B. (2010). Developing a Large-Scale Lexicon for a Less-Resourced Language : General Methodology and Preliminary Experiments on Sorani Kurdish. In *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010)*, La Valette, Malte.

Walther G., Sagot B. & Fort K. (2010). Fast Development of Basic NLP Tools : Towards a Lexicon and a POS Tagger for Kurmanji Kurdish. In *Proceedings of the 29th International Conference on Lexis and Grammar*, Belgrade, Serbie.