TALN 2010, Montréal, 19–23 juillet 2010

Développement de ressources pour le persan: lexique morphologique et chaîne de traitements de surface

Benoît Sagot[1] & Géraldine Walther[2]
(1) Alpage, INRIA Paris–Rocquencourt & Université Paris 7, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
(2) LLF, Université Paris 7, 30 rue du Château des Rentiers, 75013 Paris, France
benoit.sagot@inria.fr, geraldine.walther@linguist.jussieu.fr

Résumé. Nous présentons PerLex, un lexique morphologique du persan à large couverture et librement disponible, accompagné d'une chaîne de traitements de surface pour cette langue. Nous décrivons quelques caractéristiques de la morphologie du persan, et la façon dont nous l'avons représentée dans le formalisme lexical Alexina, sur lequel repose PerLex. Nous insistons sur la méthodologie que nous avons employée pour construire les entrées lexicales à partir de diverses sources, ainsi que sur les problèmes liés à la normalisation typographique. Le lexique obtenu a une couverture satisfaisante sur un corpus de référence, et devrait donc constituer un bon point de départ pour le développement d'un lexique syntaxique du persan.

Abstract. We introduce PerLex, a large-coverage and freely-available morphological lexicon for the Persian language, as well as a corresponding surface processing chain. We describe the main features of the Persian morphology, and the way we have represented it within the Alexina formalism, on which PerLex is based. We focus on the methodology we used for constructing lexical entries from various sources, as well as on the problems related to typographic normalisation. The resulting lexicon shows a satisfying coverage on a reference corpus and should therefore be a good starting point for developing a syntactic lexicon for the Persian language.

Mots-clés : Lexique morphologique, Persan, Développement de lexiques, Traitements de surface.

Keywords: Morphological lexicon, Persian language, Lexical development, Surface processing.

1 Introduction

La majorité des tâches de traitement automatique des langues (TAL), comme l'analyse syntaxique et la génération de langage naturel ainsi que la plupart de ses applications tels la fouille de textes, l'extraction d'informations et la traduction automatique nécessitent, ou du moins bénéficient largement, de la disponibilité de ressources lexicales à large couverture et de traitements de surface de qualité.

Parmi les ressources lexicales, les lexiques morphologiques constituent les plus simples, mais néanmoins les plus indispensables. Ils associent à chaque *forme fléchie* (ou *forme*) un lemme et une étiquette morpho-syntaxique. Cependant, pour deux raisons au moins, ce type de ressources ne semble accessible que pour un nombre bien restreint de langues convenablement décrites. D'une part, de nombreuses langues font

preuve d'un manque important de ressources et même lorsqu'il arrive qu'il existe des ressources pour ces langues, celles-ci ont le plus souvent une couverture limitée. D'autre part, les ressources existantes ne sont pas toujours librement disponibles, et ce alors même que l'expérience montre que la libre disponibilité représente le moyen le plus rapide pour le développement de ressources de qualité.

Les traitements de surface, quant à eux, recouvrent l'ensemble des opérations que l'on peut effectuer sur un texte brut pour en extraire un maximum d'informations sans avoir recours à l'analyse syntaxique. On peut donc les mettre en œuvre en tant que tels ou en préalable à une analyse syntaxique voire sémantique. Parmi les tâches généralement considérées comme relevant des traitements de surface, on peut citer la correction orthographique, l'identification des mots composés ou la reconnaissance des nombreux types d'entités nommées au sens large (des noms de personnes aux URL en passant par les adresses et les dates). Parmi ces tâches, nombreuses sont celles qui bénéficient de la disponibilité d'informations lexicales, notamment morphologiques. À l'inverse, une chaîne de traitements de surface permet la mise en place de protocoles efficaces d'amélioration des ressources lexicales.

Dans cet article nous présentons PerLex, un lexique morphologique du persan à large couverture, librement disponible et dont une version préliminaire a déjà été présentée à la conférence LREC en mai 2010 (Sagot & Walther, 2010), ainsi qu'une version persane de la chaîne de traitements de surface SXPipe (Sagot & Boullier, 2008). Après un bref état de l'art du traitement automatique du persan (section 2), nous présentons à la section 3 quelques caractéristiques de cette langue, et notamment de sa morphologie. Nous présentons ensuite le formalisme Alexina qui sous-tend le lexique PerLex (section 4) et notre description dans ce formalisme de la morphologie du persan (section 5). Nous décrivons la méthodologie de construction de PerLex et le lexique obtenu à la section 6. Enfin, nous présentons à la section 7 le travail effectué dans SXPipe pour permettre le traitement du persan, ainsi qu'une idée du taux de couverture de PerLex[1].

2 Le traitement automatique du persan : état des lieux

Le premier projet de traitement automatique ayant eu une importance notable pour le persan, le *projet Shiraz*, a été consacré à la traduction automatique du persan vers l'anglais (Amtrup *et al.*, 2000). Il a eu notamment pour résultat la mise en place d'un lexique bilingue d'environ 50 000 entrées[2] qui s'appuie sur une description du persan au sein d'un modèle de grammaire d'unification (Megerdoomian, 2000). Ce lexique a ensuite été adapté aux outils Xerox pour les automates à états finis (Megerdoomian, 2004).

Outre les travaux réalisés dans le cadre du projet Shiraz, 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 es-

[1] Le développement de PerLex s'inscrit dans un projet plus large dénommé PerGram (Projet Franco-Allemand ANR/DFG MU 2822/3-1). Le but de PerGram est (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 PerLex en tant que lexique associé à cette grammaire et pour lequel il est prévu qu'il inclue à terme aussi bien les informations morphologiques que syntaxiques des lexèmes. PerGram est un projet dirigé conjointement par Pollet Samvelian (Université Paris 3) et Stefan Müller (Freie Universität Berlin).

[2] Ce lexique ne semble cependant pas être disponible librement.

[3] `http://sourceforge.net/projects/perstem/`

sentiellement 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 est une langue indo-européenne de la famille des langues indo-aryennes, et plus précisément du groupe des langues iraniennes occidentales. Le persan se caractérise par un ordre des mots de type SOV relativement fixe. Son nombre de locuteurs avoisine les 130 millions, répartis majoritairement en Iran, en Afghanistan, Tadjikistan et en Ouzbékistan.

3.1 Écriture et translittération

Le persan s'écrit de droite à gauche au moyen d'une variante de l'alphabet arabe, avec quelques caractères en plus, d'autres en moins, et d'autres qui ont une forme légèrement modifiée. 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.

L'encodage de ces caractères au moyen de la norme Unicode peut ainsi se faire de deux façons : soit au moyen de caractères contextuels, qui représentent la forme précise que doit prendre le caractère (lié ou non, à droite et à gauche), soit au moyen de caractères génériques, dont la forme varie en fonction du contexte. L'encodage au moyen des caractères contextuels est aujourd'hui considéré comme obsolète, car il ne respecte pas l'unicité de représentation de chaque caractère. L'encodage au moyen des caractères génériques nécessite cependant parfois l'introduction d'*espaces de longueur nulle* (*zero-width non-joiner*, ou *ZWNJ*) entre deux caractères pour indiquer qu'ils ne doivent pas être liés alors qu'ils pourraient l'être. En effet, en persan, certains affixes s'écrivent collés mais non liés au morphème auquel ils se rattachent, quand bien même ce rattachement rapproche deux caractères qui seraient autrement liés.

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, dont une variante ne fait usage que de caractères que l'on peut représenter au moyen de l'encodage ISO-8859-2 (dit *Latin-2*). L'avantage est que l'on peut alors utiliser sans problème des outils qui ne gèrent que les représentations *8-bits*, c'est-à-dire les représentation où les notions informatiques et typographiques de caractère coïncident[4].

En plus d'un outil de translittération permettant de basculer des caractères persans aux caractères latins correspondants dans un sens ou dans l'autre, nous avons développé deux outils de normalisation. Le premier remplace tout caractère contextuel par le caractère générique correspondant, en insérant si besoin des espaces de longueur nulle. Le second remplace tout caractère qui ne fait pas partie de l'alphabet persan *stricto sensu* par sa contrepartie appropriée.

[4]Pour plus de lisibilité, nous employons ici une transcription phonétique plus standard comprenant aussi les voyelles brèves.

3.2 Aperçu de la morphologie du persan

La morphologie nominale du persan n'affiche qu'un nombre restreint de formes fléchies. Il n'y a pas de flexion casuelle, pas de différence de genres (mis à part pour quelques emprunts arabes animés qui prennent la marque arabe du féminin en *-e*). Le persan n'a que deux nombres, le singulier et le pluriel, dont seul le pluriel se distingue par une marque morphologique spécifique, soit le suffixe *-hâ* (possible pour tous les noms comptables), soit, pour certains noms désignant principalement des animés, le suffixe *-ân*, plus formel. S'ajoutent à cela quelques marques de pluriel arabes, *-ât*, *-un*, *-in* etc., qui ne s'attachent qu'à des emprunts arabes. Le persan a également hérité d'un certain nombre de pluriels dits *brisés*, mais ces derniers ne s'analysent plus directement en morphologie. Par ailleurs, il existe également une particule enclitique spécifique, appelée *Ézafé*, qui fonctionne comme un marqueur de dépendance. Il peut marquer aussi bien les noms que les SN en tant qu'éléments modifiés (Samvelian, 2007; Lazard *et al.*, 2006).

Par ailleurs, le persan dispose d'un déterminant indéfini enclitique *-i*, sans formes séparées pour le singulier et le pluriel. Si cette particule se combine avec un nom modifié par un adjectif, il peut soit s'attacher directement au nom, soit suivre l'adjectif. Les autres particules enclitiques du persan sont le *-i* qui se combine avec le marqueur relatif *ke*, le marqueur optionnel de définitude *-(h)e* et la postposition *-râ* indiquant l'objet direct défini. Les adjectifs, quant à eux, ne varient qu'en degré, prenant le suffixe *-tar* au comparatif et le suffixe *-tarin* au superlatif. Ils peuvent néanmoins être suivi de l'Ézafé lorsqu'ils suivent un nom modifié plus précisément ou lorsqu'ils prennent eux-mêmes un objet direct ou indirect. Ce dernier point s'applique en particulier au adjectifs dérivés de verbes (Lazard *et al.*, 2006).

Les verbes simples du persan sont en nombre très réduit. Ils constituent une classe fermée de seulement 200 unités environ, la majorité des sens pris en charge par des verbes dans la plupart des langues du monde étant ici exprimés par des locutions verbales complexes qui constituent un procédé très productif. La morphologie verbale du persan est légèrement plus complexe que sa morphologie nominale, mais elle suit néanmoins un schéma assez simple. On admet habituellement (Lazard *et al.*, 2006) l'existence de deux radicaux par verbe, l'un servant à la formation des temps du présent, du participe présent, du gérondif et de l'impératif, l'autre aux temps du passé ainsi qu'aux participes passé et d'obligation/de possibilité et aux infinitifs. Les temps composés et la voie passive sont dérivés à partir du participe passé. Les paradigmes verbaux du persan s'appuient sur un modèle de type *Préfixe(s)-Rad-Suffixe(s)*.

Dans ce modèle, les préfixes possibles sont les deux préfixes TAM[5] *mi-* et *be-*, les suffixes les désinences personnelles *-am*, *-i*, *-ad/-e/ø*, *-im*, *-id/-in* et *-and/-an*. Les combinaisons que l'on peut ainsi obtenir génèrent sept temps/modes verbaux différents pour six personnes chacun, ainsi que cinq formes verbales nominales pouvant à leur tour se combiner avec les particules enclitiques citées plus haut. Les préfixes de négation *n-* ou *m-* (formel pour les formes de l'impératif notamment) peuvent également se combiner avec les formes verbales ainsi générées (Lazard *et al.*, 2006). Par ailleurs, on note également que le persan possède deux paradigmes de l'auxiliaire *être*, l'un étant constitué de mots pleins, l'autre de formes enclitiques qui peuvent venir s'attacher à des noms ou des adjectif, mais aussi à la forme du participe passé avec lequel elles forment par ce fait les paradigmes du parfait et de l'imparfait composé (Lazard *et al.*, 2006).

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

[5]Temps-Aspect-Mode.

4 Le formalisme lexical Alexina

Nous avons développé le lexique morphologique PerLex dans le formalisme Alexina. Ce formalisme couvre à la fois le niveau morphologique et le niveau syntaxique (p.ex. la valence), ce qui sera utile pour le développement futur de PerLex [6]. 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). Un certain nombre de ressources lexicales existe dans ce formalisme, telles que le Le*fff*, un lexique morphologique et syntaxique à large couverture pour le français (Sagot *et al.*, 2006; Sagot, 2010), ainsi que des ressources pour l'espagnol (le Le*ff*e), le galicien, le polonais (Sagot, 2007), le slovaque (Sagot, 2005), et bientôt l'anglais, et aussi SoraLex, un lexique morphologique pour le kurde sorani (Walther & Sagot, 2010), une langue typologiquement proche du Persan s'écrivant avec une variante de l'alphabet arabe.

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.

5 Formalisation de la morphologie du persan

Parmi les diverses particules enclitiques citées à la section 3.2, toutes n'ont pas été retenues pour être traitées en morphologie. PerLex suit les choix linguistiques adoptés au sein du projet PerGram, définis notamment dans (Samvelian, 2007) :
– les marques de pluriel, l'Ézafé *-(y)e* (dès lors qu'il est représenté à l'écrit), le marqueur indéfini *-i*, le comparatif et le superlatif *-tar* et *-tarin*, les suffixes personnels (les désinences verbales) et les formes enclitiques de l'auxiliaire *être* sont considérés comme des suffixes flexionnels,
– dans sa combinaison avec le marqueur relatif *ke*, la particule enclitique *-i* est considérée comme un mot composé amalgamé au mot précédent,
– les autres enclitiques, y compris la copule lorsqu'elle ne sert pas à former les temps du parfait et de l'imparfait composé, et le *-râ* marqueur de l'objet direct défini, sont interprêtés comme des formes indépendantes quoique amalgamées.

En ayant en tête ces choix linguistiques, nous avons développé une description complète de la morphologie du Persan dans le formalisme morphologique d'Alexina (Sagot, 2007), à partir des données de (Lazard *et al.*, 2006). Dans ce formalisme, la flexion est modélisée comme l'affixation d'un préfixe et d'un suffixe de part et d'autre d'un radical. Des phénomènes de *sandhi*[7] peuvent avoir lieu aux frontières de ces unités, qui sont éventuellement conditionnés par des propriétés du radical.

Notre description contient entre autres 27 tables verbales et 5 tables nominales.

[6] Dans cet article, nous ne prenons pas en compte les informations syntaxiques, puisque notre travail concerne la construction d'un lexique morphologique.

[7] Transformations sur le radical et/ou l'affixe provoquées par la juxtaposition ce ces derniers. Ainsi en français, lorsque le suffixe *-ons* est juxtaposé au radical *mang-*, un phénomène de *sandhi* a lieu qui induit l'insertion d'un *e*.

6 Construction du lexique PerLex

Nous avons exploité différentes sources d'informations lexicales librement disponibles, avec une importance respective qui varie d'une partie du discours à l'autre :
– le corpus BijanKhan (BijanKhan, 2004; Amiri *et al.*, 2007), un corpus annoté automatiquement en parties du discours ;
– la Wikipedia Persane[8],
– un lexique nominal du Persan en cours de construction par Mehdi Ghassemi à l'Université Paris-Est (Ghassemi, c.p.),
– la grammaire de référence de Lazard *et al.* (2006),
– l'introspection par des linguistes locuteurs natifs du Persan.

Les entrées lexicales ont été créées en trois étapes :

1. construction d'entrées lexicales à partir des ressources ci-dessus ;
2. nettoyage des entrées obtenues ;
3. ajout manuel d'entrées manquantes à partir des formes trouvées dans le corpus BijanKhan mais non couvertes par les entrées déjà construites.

Construction du lexique de base Dans un premier temps, nous avons développé et vérifié manuellement une liste de lemmes verbaux (à l'infinitif) à partir de conjugueurs en ligne librement disponibles sur internet. Nous avons associé à chacun de ces lemmes une des classes flexionnelles de notre description morphologique (parfois plusieurs).

La Wikipedia persane a été exploitée de la façon suivante. Nous avons fait un inventaire des *Catégories* Wikipedia indiquant un article concernant une personne ou une ville. En récupérant et normalisant les titres de chaque article concerné, ainsi que les titres de tous les articles redirigeant vers ce même article, nous avons pu constituer un lexique de noms de personnes et de noms de villes. Nous avons complété ces données par la liste des pays telle que disponible à l'article correspondant. Le résultat est un ensemble de plus de 10 000 lemmes pour des noms propres, auxquels nous avons attribué une classe de flexion nominale ne permettant pas la mise au pluriel.

Par ailleurs, une liste de lemmes nominaux a été extraite à partir de données de Mehdi Ghassemi. Pour certains d'entre eux, la forme de base du pluriel était mentionnée. Pour les autres, une recherche dans le corpus BijanKhan a permis d'assigner une ou deux formes de base pour le pluriel.

Enfin, nous avons passé le corpus BijanKhan à travers notre normaliseur, et nous avons extrait du résultat d'autres entrées nominales, ainsi que des entrées pour les autres catégories[9]. Ces entrées, notamment pour les classes fermées, ont été complétées à la main à partir de (Lazard *et al.*, 2006). Nous avons attribué à toutes les entrées adjectivales l'unique classe flexionnelle de cette catégorie. Pour les catégories restantes, toutes les entrées ont été considérées comme invariables, à l'exception de certains adverbes qui peuvent recevoir le marqueur indéfini *-i*, l'Ézafé ou les affixes personnels[10].

[8]La Wikipedia persane est accesible en ligne à l'adresse `http://fa.wikipedia.org`. Nous avons exploité l'extraction au format *wiki* en date du 16 février 2010.

[9]Pour les lemmes auxquels le corpus BijanKhan associe plusieurs catégories, nous avons éliminé automatiquement les entrées correspondant à des catégories concernant moins de 1% de ses occurrences. Cela permet de réduire le bruit qui provient des erreurs d'annotation du corpus.

[10]En Persan, quasiment tout adjectif peut être employé comme adverbe. Par conséquent, les entrées adverbiales qui correspondaient à une entrée adjectivale pour le même lemme ont été éliminées.

Nettoyage du lexique de base À ce stade, et malgré le filtrage évoqué à la note 9, de nombreuses entrées lexicales sont incorrectes, notamment une proportion significative de celles extraites du corpus BijanKhan. C'est là une conséquence immédiate du fait que ce dernier a été annoté automatiquement, avec un taux d'erreur non nul. Nous avons donc supprimé un certain nombre d'entrées lexicales, et notamment toutes les entrées nominales et adjectivales extraites du corpus BijanKhan correspondant à des lemmes qui avaient toutes les caractéristiques de surface d'un pluriel (suffixe *-hâ*, notamment, suivi ou non de l'Ézafé, du suffixe de l'indéfini ou des suffixes personnels). Nous avons également éliminé les nombreuses entrées correspondant manifestement à des erreurs typographiques, en particulier lorsqu'un *espace de longueur nulle* a été supprimé, ou à l'inverse remplacé par un espace (ainsi les entrées erronnées pour les pronoms *ân hâ* et *ânhâ* en plus du correct *ân_hâ*[11]). Enfin, nous avons supprimé tous les caractères jugés superflus, bien que correctement pris en compte par nos outils (signes de vocalisation, etc.). Le résultat de ce lexique de base constitue une première version de lexique morphologique du persan.

Extension du lexique de base Une fois cette première version de PerLex construite, nous avons cherché les formes attestées dans le corpus mais non couvertes par notre lexique. Naturellement, ces formes peuvent correspondre à des entrées manquantes, à des entrées incorrectes, à des erreurs orthographiques ou à des erreurs typographiques. Nous les avons donc classées par fréquence d'apparition, et avons complété manuellement le lexique à partir de ces données. Les entrées manquantes étaient principalement des noms propres appartenant à des catégories différentes de celles cherchées dans Wikipedia (noms de continents, noms de régions d'Iran, etc.), ainsi qu'à des pluriels brisés. Nous avons cependant constaté que de très nombreuses formes inconnues étaient liées à la typographie, sans que cela n'ait été détecté pendant la phase de nettoyage.

Le lexique obtenu Le lexique obtenu contient 35 914 entrées intensionnelles (de niveau lemme) qui produisent 524 700 entrées extensionnelles (de niveau forme) pour 494 488 formes distinctes. Quelques informations complémentaires sur la distribution par partie du discours sont fournies à la table 1.

Partie du discours	entrées intensionnelles	lemmes distincts	entrées extensionnelles
verbes	171	139	19 776
noms communs	9 553	9 106	177 988
noms propres	10 996	10 938	33 076
adjectifs	11 872	11 835	290 537
autres	3 322	3 120	3 323
total	*35 914*	*33 454*	*524 700*

TAB. 1 – Données quantitatives sur PerLex

7 Traitements de surface pour le persan

SXPipe Dans (Sagot & Boullier, 2008), les auteurs présentent SXPipe, une chaîne multilingue de traitements de surface qui prend en entrée des textes bruts. SXPipe, originellement développé pour le français, a été étendu depuis à d'autres langues. Avant les travaux rapportés ici, et outre le français, SXPipe traitait à des degrés divers l'espagnol, l'anglais, le néerlandais, le polonais, le slovaque et l'italien. SXPipe

[11] Le blanc souligné (_) indique un *espace de longueur nulle*.

est utilisé pour le français en préliminaire à divers analyseurs syntaxiques, mais également pour diverses langues en tant que tel, par exemple pour la détection d'entités nommées, la correction et normalisation de textes ou la détection et l'attribution de citations verbatim.

Le traitement d'un corpus par SXPipe peut se décomposer en six étapes principales : (1) normalisation typographique du corpus (p.ex. gestion des encodages) ; (2) reconnaissance de certaines entités nommées de surface dans le texte brut (URL, dates, nombres en chiffre, adresses...) ; (3) découpage du texte brut en tokens et en phrases ; (4) reconnaissance de certaines entités nommées au niveau des tokens (séquences en langue étrangère...) ; (5) regroupement non-déterministe des tokens en formes (correction orthographique et traitement des composés et des amalgames), qui résulte en un graphe de formes (outil TEXT2DAG) ; (6) reconnaisance de certaines entités nommées au niveau du graphe de formes (nombres en toutes lettres, noms de personnes, de lieux, d'organisations...).

Traitement du persan dans SXPipe Jusqu'à présent, SXPipe ne gérait que des langues utilisant des variantes de l'alphabet latin. Nous avons donc adapté SXPipe, notamment au niveau de l'étape 1 et en rajoutant une étape à la fin de la chaîne, pour qu'il soit en mesure de traiter les langues utilisant des variantes de l'alphabet arabe. En particulier, nous avons ajouté au début de l'étape 1 les outils de normalisation évoqués à la section 3.1. Nous avons également inséré dans la chaîne notre outil de translittération afin de permettre aux modules qui en ont besoin de traiter un texte au format 8-bit. C'est notamment le cas de TEXT2DAG, dont nous avons compilé une version pour le persan à partir de PerLex et de règles (pondérées) de correction orthographique dont certaines sont spécifiquement adaptées à cette langue.[12] En revanche, aucun travail d'adaptation des différents modules de reconnaissance des entités nommées n'a encore été effectué. Seuls les nombres écrits en chiffres sont reconnus, par le simple fait que la translittération de nombres écrits en persan donne précisément des nombres similaires à ceux rencontrés en français, qui sont donc reconnus par le module correspondant.

Évaluer la couverture de PerLex L'outil de correction orthographique et de reconnaissance des amalgames et des composés ainsi construit pour le persan permet de faire la différence entre un token qui est un vrai "mot inconnu" et un token qui ne correspond pas à une forme connue du lexique mais que l'on peut analyser comme des amalgames de formes connues du lexique. Ainsi, nous avons pu évaluer le taux de couverture de PerLex sur un corpus de taille importante. Nous avons pour cela converti le corpus BijanKhan original en un corpus « brut », en reconstituant les phrases à partir des mots qui les composent, et en recollant les signes de ponctuation aux mots qui les précèdent. Nous avons ensuite essayé de distinguer les tokens analysables, les tokens contenant des imperfections typographiques résiduelles et les tokens correspondant vraiment à des mots inconnus. L'estimation que nous obtenons est la suivante : 97,0% de couverture (96,2% de tokens analysables et 0,8% de tokens corrigibles en tokens analysables par normalisation typographique), et 3% de tokens hors couverture.

8 Conclusion et perspectives

Nous avons présenté la première version de PerLex, un lexique morphologique du persan à large couverture, ainsi qu'une version persane de la chaîne de traitements de surface SXPipe. PerLex est distribué sous

[12] Aucune évaluation de ce correcteur orthographique n'a pu être effectuée pour l'instant.

license libre LGPL-LR sur la page internet du projet Alexina[13]. SXPipe est distribué sous license libre LGPL, en tant que sous-projet de la Chaîne Linguistique Alpage, par exemple sur la page de ce projet[14].

Pour l'instant, le lexique PerLex est uniquement morphologique et nécessite une validation manuelle complète, qui est actuellement en cours au sein du projet franco-allemand ANR/DFG PerGram. En parallèle, l'ajout des informations syntaxiques (y compris les cadres de sous-catégorisation) est en cours sur les entrées pour les verbes simples. Les autres catégories devront voir également leurs entrées complétées par des informations syntaxiques. Par ailleurs, PerLex sera étendu à la description du phénomène des prédicats verbaux complexes, important en Persan. Ces deux tâches seront réalisées en tout ou partie au moyen de techniques automatiques déjà utilisées pour le développement d'autres lexiques Alexina (Nicolas *et al.*, 2008), puis suivies d'étapes de validation manuelle.

Quant à la chaîne de traitements de surface SXPipe, elle n'effectue sur le persan qu'une partie des tâches disponibles pour d'autres langues, à savoir la normalisation typographique, la romanisation (si demandée), la reconnaissance d'un nombre très limité de types d'entités nommées et la correction orthographique et reconnaissance des mots composés. Les entités nommées plus standard (noms de personnes, de lieux, d'organisations, mais également dates, adresses, etc.) devront être gérées. Mais SXPipe constitue déjà une chaîne viable, que nous comptons compléter par l'étiqueteur morphosyntaxique MElt (Denis & Sagot, 2009) entraîné sur le corpus BijanKhan et couplé avec PerLex. Ceci permettra d'améliorer le lexique morphologique PerLex au moyen de techniques telles que décrites par exemple dans (Nicolas *et al.*, 2008).

Références

AMIRI H., HOJJAT H. & OROUMCHIAN F. (2007). Investigation on a Feasible Corpus for Persian POS tagging. In *Proceedings of the 12th International CSI Computer Conference (CSICC)*.

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.

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., MANDY PET & SORIA C. (2006). Lexical Markup Framework (LMF). In *Proceedings of LREC'06*, Gênes, Italie.

[13] `http://alexina.gforge.inria.fr`
[14] `http://lingwb.gforge.inria.fr`

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, Mexique.

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

NICOLAS L., SAGOT B., MOLINERO M. A., FARRÉ J. & VILLEMONTE DE LA CLERGERIE É. (2008). Computer aided correction and extension of a syntactic wide-coverage lexicon. In *Proceedings of COLING'08*, Manchester, Royaume-Uni.

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, Proceedings of TSD'05*, p. 156–163, Karlovy Vary, République Tchèque : Springer-Verlag.

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

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

SAGOT B. & BOULLIER P. (2008). SXPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, **49**(2), 155–188.

SAGOT B., CLÉMENT L., DE LA CLERGERIE E. & BOULLIER P. (2006). The Le*fff* 2 syntactic lexicon for French : architecture, acquisition, use. In *Proceedings of the 5th Language Resource and Evaluation Conference*, Lisbonne, Portugal.

SAGOT B. & WALTHER G. (2010). A Morphological Lexicon for the Persian Language. In *Proceedings of LREC 2010)*, La Valette, Malte. À paraître.

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

SHAMSFARD M. & FADAEE H. (2008). A Hybrid Morphology-Based POS Tagger for Persian. In *Proceedings of the 6th Language Resources and Evaluation Conference*, 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 Workshop)*, La Valette, Malte. À paraître.