A new morphological lexicon and a POS tagger for the Persian Language

Benoît Sagot[1], Géraldine Walther[2,3], Pegah Faghiri[3], Pollet Samvelian[3]

1. Alpage, INRIA Paris–Rocquencourt & Université Paris 7, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
2. Laboratoire de Linguistique Formelle, CNRS & Université Paris 7, 175 rue du Chevaleret, 75013 Paris, France
3. UMR 7528 Mondes iranien et indien, CNRS & Université 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

In (Sagot and Walther, 2010), the authors introduce an advanced tokenizer and a morphological lexicon for the Persian language named PerLex. In this paper, we describe experiments dedicated to enriching this lexicon and using it for building a POS tagger for Persian.

Natural Language Processing (NLP) tasks such as part-of-speech (POS) tagging or parsing as well as most NLP applications require large-scale lexical resources. Yet, such resources rarely are freely available, even though it is the fastest way to building high-quality resources. In this paper, we introduce a new version of the large-scale and freely available morphological lexicon for Persian named PerLex, which relies on a new linguistically motivated POS inventory as well as several validation steps; we show how we used this new lexicon for generating an improved version of the BijanKhan corpus (BijanKhan, 2004) and training the MElt tagging system (Denis and Sagot, 2009), thus creating a freely available Persian tagger.

The first important NLP project on Persian is the Shiraz project, targeted towards Persian to English automatic translation (Amtrup *et al.*, 2000). Among other things, it produced 50,000 terms bilingual lexicon (which however does not seem to be freely available) based in part on a unification-based description of the Persian morphology (Megerdoomian, 2000). Apart from the Shiraz project, some other NLP tools such as morphological tools and lemmatisers have been developed, although not associated with a full large scale lexicon (cf. the freely available lemmatizer PerStem (Dehdari and Lonsdale, 2008)). To our best knowledge, the only freely available large-coverage lexical resources for Persian are the above-mentioned PerLex lexicon (Sagot and Walther, 2010) and the Persian lexicon within MULTEXT-East version 4 (Erjavec, 2010; QasemiZadeh and Rahimi, 2006). Other recent work on the development of NLP tools and resources for Persian processing is mostly focused on designing part-of-speech taggers (QasemiZadeh and Rahimi, 2006; Shamsfard and Fadaee, 2008), parsers (Dehdari and Lonsdale, 2008) or automatic translation systems.

Improving PerLex The PerLex 1 lexicon (Sagot and Walther, 2010) contained approx. 36,000 lexical entries (lemmas) corresponding to over 520,000 (inflected) form entries describing approx. 500,000 unique forms. Apart from its underlying morphological description, PerLex 1 had mainly been built automatically using automatic lexical data aquisition techniques such as the extraction of lexical entries from the automatically tagged BijanKhan corpus (BijanKhan, 2004) and from Wikipedia. Therefore, the first step towards the construction of a new version, PerLex 2, was to improve the quality of the lexicon by validating the entries extracted from the BijanKhan corpus. We first automatically (pre-)validated a certain amount of entries, using comparison and/or fusion of PerLex with other lexical resources (i.e. the Persian lexicon included in version 4 of MULTEXT-East (henceforth MTE4-fa) (QasemiZadeh and Rahimi, 2006; Erjavec, 2010) and the Persian Pronunciation Dictionary (henceforth PPD) (Deyhime, 2000). Being not freely distributable, we didn't use the PPD to provide us with additional entries, but only to pre-validate existing lexical entries, in particular those for which most inflected forms are found in the PPD. On the other hand, MTE4-fa is a freely available and redistributable morphological lexicon including 13,006 lexical entries. We established a mapping between POS tags found in MTE4-fa and in PerLex, converted MTE4-fa in same format as PerLex and merged it with PerLex. The entries resulting from merging entries from both resources were considered pre-validated. Entries corresponding only to MTE4-fa entries were added to PerLex (in many cases, this required to add the appropriate inflection class manually).

Entries automatically pre-validated were excluded from the manual validation (apart for nouns and adjectives) hence avoiding unecessary manual validation costs. So far, we have carried out two seperate manual validation campaigns using a dedicated online validation interface that aims at optimizing validation speed (for example, lexical entries are displayed as a canonical form and the minimal set of inflected forms whose correctness guarantees that the entry's inflection class is correct; another example is that the interface allows for specifying most types of inflection class assignment errors (e.g., a lemma ending in ی *yeh* pronounced [i] but considered as if it was pronounced [j]). The first validation campaign created 751 validation tickets (451 correct entries, 250 correct POS but invalid inflected forms, no invalid POS and 50 completely invalid entries, mostly due to encoding bugs we resolved in the meantime). The second validation campaign created 1.097 validation tickets (818 correct entries, 17 valid POS but invalid inflected forms, 26 invalid categories ans 129 completely invalid entries, mostly inflected pronominal forms erroneously considered as individual lexical entries).

Another new feature of PerLex 2 is its new sound set of POS. PerLex 1 had simply adopted the POS used in the BijanKhan corpus (BijanKhan, 2004; Amiri *et al.*, 2007). We decided to convert the lexicon into a new set of linguistically motivated POS (Faghiri & Samvelian, in prep.): nouns, proper nouns, adjectives, adverbs; verbs, prepositions, conjunctions, classifiers, pronouns, determiners and interjections. The conversion has been realised through automatic conversion techniques. It was straightforward for nouns (`N`), verbs (`V`), proper nouns (`PN`), pronouns (`PRO`), interjections (`INT`), delimiters (`DELM`). For the other POS, precise criteria had to be established manually to re-assign their members. The POS `MORP` of the BijanKhan corpus has been altogether suppressed since it contained elements contributing to word-formation in various ways but not considered words in the description we adopted. On the other hand, we established a new POS tag for classifiers (`CLASS`) which replaces the old specifier-tag `SPEC`.

The size of PerLex 2 is similar to that of PerLex 1 (suppressing erroneous entries has quantitatively counter-balanced the addition of new entries and the conversion into a new POS set does not result in quantitative differences), yet it is the qualitative improvement, such as the addition of new inflection tables for auxiliairies and light verbs, that characterises PerLex 2.

Corpus modification The next step of our work was to develop a new tagger for Persian based on our POS inventory and on PerLex 2, using the MElt tagging system (Denis and Sagot, 2009). We first designed a tagset that is a refinment of this inventory. Our tagset defines 79 tags, among which 37 verbal tags, 9 pronominal tags and 8 nominal tags.

For training the MElt system, we decided to create a new version of the BijanKhan corpus. This new version differs in two ways: first, we improved the original automatic tokenization and annotation of the corpus. Second, we converted the corpus so that it uses our tagset. We started from the version of the corpus used in (Sagot and Walther, 2010), which is already segmented in 88,885 sentences. We applied rule-based transformations for correcting systematic tokenization and/or annotation errors. These include among others various kinds of typographic (e.g., whitespace) inconsistencies (verbal prefixes, nominal suffixes, acronyms, compound prepositions, removal of the `MORP` category, and others), whose correction require modifications in the annotation itself. We also corrected systematic annotation errors. Next, we needed to convert the corpus annotations into our 79-tag tagset. In order to achieve a good level of quality, we decided to convert mostly the annotation of those tokens for which we could find a unique tag from our tagset that was consistent with both the corrected corpus annotation and lexical information in PerLex 2. However, in rare cases, heuristics allowed us to choose among various possible tags, as well as to convert annotations for tokens unknown to PerLex (e.g., by relying on morphology-based patterns). The resulting modified BijanKhan corpus was then split in 3 parts. The last 100 sentences (1,568 of their 1,707 tokens could be converted) were extracted and the annotations manually converted (when needed) or corrected, leading to a *gold standard*. Among the remaining sentences, those for which all tokens had been successfully converted constitute a 18,731-sentence *training corpus* (302,690 tokens).

Tagging Persian with MElt$_{\text{fa}}$ Next, we extracted from PerLex 2 a lexicon based on our 79-tag tagset. Together with the above-described (far from error-free) training corpus, this allowed us to train the MElt system and generate a tagger for Persian, MElt$_{\text{fa}}$. W.r.t. our gold standard, MElt$_{\text{fa}}$ has a 90.3% accuracy on the full tagset, and a 93.3% accuracy if we project this tagset on our 14 POS inventory. Evaluated only on the 1,568 tokens for which the annotations could be converted automatically, these figures reach respectively 93.9% and 95.3%. These figures are probably a lower bound on the accuracy we would reach if all annotations were converted successfully. Indeed, non-converted tokens have not been converted in the training data either: MElt has not learned any contextual information about them, hence more errors on these tokens (this in turn might affect MElt$_{\text{fa}}$'s decisions on surrounding tokens).

We compared the quality of MElt$_{\text{fa}}$'s annotations to those resulting from our automatic conversion process. It turns out that the accuracy of these annotations on those 1,568 tokens for which the automatic conversion was successful is exactly the same (93.9% and 95.3%) as that of MElt$_{\text{fa}}$, although only 48% concern the same tokens. In other words, on these 1,568 tokens, MElt$_{\text{fa}}$ was able to produce annotations whose quality is the same as the quality of its training corpus, which in turn is higher than that of the original BijanKhan corpus. We believe that this is related both to the use of PerLex as a source of information and to the fact that MElt$_{\text{fa}}$'s probabilistic model smoothes many errors in its training corpus (with a ``co-training''-like effect). This latter hypothesis is confirmed by the fact that, among these 1,568 tokens, MElt$_{\text{fa}}$'s result are slightly closer to the gold standard (93,9% accuracy on the full tagset) than to its automatically converted version before manual correction (93.4%).

References

Amiri, H., H. Hojjat, and F. Oroumchian. 2007. Investigation on a feasible corpus for Persian POS tagging. In *Proceedings of the 12th International CSI Computer Conference (CSICC).*

Amtrup, Jan W., Hamid Mansouri Rad, Karine Megerdoomian, and Rémi Zajac. 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, Jon and Deryle Lonsdale. 2008. A Link Grammar Parser for Persian. In S. Karimi, V. Samiian and D. Stilo, Eds., *Aspects of Iranian Linguistics*, volume 1. Cambridge Scholars Press.

Denis, Pascal and Benoît Sagot. 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 Pron. Dict.).* Tehran, Iran: Farhang Moaser Publishers.

Erjavec, Tomaž. 2010. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of LREC'10*, Valetta, Malta.

Megerdoomian, Karine. 2000. Unification-based Persian morphology. In *Proceedings of CICLing 2000*, Mexico.

QasemiZadeh, Behrang and Saeed Rahimi. 2006. Persian in Multext-East Framework. In *FinTAL*, p. 541--551.

Sagot, Benoît and Géraldine Walther. 2010. A morphological lexicon for the persian language. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta: ELDA.

Shamsfard, Mehrnoush and Hakimeh Fadaee. 2008. A hybrid morphology-based pos tagger for Persian. In N. Calzolari, Ed., *Proceedings of LREC'08*, Marrakech, Morocco.