A Morphological Lexicon for the Persian Language

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

Abstract

We introduce PerLex, a large-coverage and freely-available morphological lexicon for the Persian language. 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 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.

1. Introduction

Most Natural Language Processing (NLP) tasks such as part-of-speech tagging, shallow or deep parsing and natural language generation, as well as most NLP applications such as data mining, information extraction, or automatic translation require or strongly benefit from the availability of large-scale lexical resources. Among them, morphological lexicons are the most basic yet the most needed resources. They associate a lemma and a morphosyntactic tag with each known wordform (form, in short). However, for at least two reasons, such resources are virtually available only for a restricted set of well-described languages. First, many languages are resource-scarce, and existing lexical resources for these languages, if any, do not have a large coverage. Second, existing lexical resources are not always freely available, although experience shows that free availability is the fastest way to high-quality resources.

In this paper, we introduce PerLex, a new large-scale and freely available morphological lexicon for the Persian language. We briefly describe the Persian language and Persian morphology, the Alexina framework used for our lexical development work, the origin of the lexical data in PerLex, and the resulting lexical resource.

2. Related Work

The first NLP project of importance on Persian language is the Shiraz project, targeted towards Persian to English automatic translation (Amtrup et al., 2000). It resulted, among others, in the construction of a 50,000 terms bilingual lexicon,[1] based in part on a unification-based description of the Persian morphology (Megerdoomian, 2000), later redesigned for using the Xerox finite-state machinery (Megerdoomian, 2004).

Apart from the work related to the Shiraz project, some other NLP tools, such as morphological tools and lemmatisers, have been developed. Yet they have not led to the construction of a large-scale lexicon. Among those works are those of (2008), especially their freely available lemmatiser PerStem.[2]

Other recent work in the development of NLP tools and resources for Persian processing is mostly focused on designing part-of-speech taggers (QasemiZadeh and Rahimi, 2006; Tasharofi et al., 2007; Shamsfard and Fadaee, 2008), parsers (Hafezi, 2004; Dehdari and Lonsdale, 2008) or automatic translation systems (Feili and Ghassem-Sani, 2004; Saedi et al., 2009).

3. The Persian Language

The Persian language is a inflectional SOV language with a relatively fixed word order, belonging to the Western Iranian branch of Indo-European languages. It is spoken by around 130 million people, mainly in Iran, Afghanistan and Tajikistan and Uzbekistan but also in Pakistan, Bahrain, Iraq, Kazakhstan and the Iranian diaspora. In Iran, where Persian functions as the official language, it is also often referred to as Dari, Farsi or Parsi.

3.1. Persian Script and Transliteration

Persian is written from right to left and uses a modified version of the Arabic script. Some characters have been added, others are not used and some show a slightly modified shape. As in Arabic, the only diacritic representation of short vowels is usually not written, nor is there a difference between capital and lower case letters. Moreover, two adjacent characters may either be joined (that is written with one uninterrupted line, possible only with some letters) or juxtaposed (that is following each other directly yet without being joined) or separated by a white space. Depending on whether a given character is joined to or isolated from its adjacent characters, this same character can adopt up to three different shapes.

Therefore, encoding these characters within Unicode can be done in two different ways: either one directly uses the appropriate contextual character which represents the exact shape a given character has to adopt in that one given context (left- or right-joined or not), or one resorts to the use of generic characters whose shape automatically varies according to the context. Using contextual characters has however become obsolete nowadays since they do not allow for the representation letters by one unique character each. Using the generic character therefore sometimes requires the insertion of zero-width non-joiners (*ZWNJ*) between

[1] This lexicon doesn't seem to be freely available.

[2] http://sourceforge.net/projects/perstem/

two characters to indicate that in a particular case they must not be joined whereas otherwise they could. Some Persian affixes for example have to be written without being separated by a white space but must yet not be joined to the preceding morpheme even if the letters they are represented by normally would.
Some of the resources we developed during on our work is based on a transliteration of Persian into Latin characters. We have adopted the bijective transliteration system which has been developed within the PerGram project and which has a version solely based on characters that can be found within the ISO-8859-2 (or Latin-2) encoding. This allows for an effortless use of tools that internally require 8-bit encodings, i.e., encodings for which the typographic and electronic notions of character coincide.[3]

3.2. Persian Morphology in Brief

Within Persian morphology, the nominal system is relatively simple. There is no case system, nor gender distinction (except for a few animate Arabic loanwords which take the feminine ending *-e*). Persian shows but two number features, singular and plural, of which only the plural is marked by either the suffix *-hâ* (which works for all countable nouns) or, for some animated nouns only, the more formal suffix *-ân*, or one of the Arabic plural markers *-ât*, *-un*, *-in* etc. attaching solely to Arabic loanwords (Lazard et al., 2006). Persian also possesses a few broken plurals directly inherited with Arabic loanwords. But those plurals are no longer treated within morphology. Further, there exists a specific enclitic particle *-(y)e* for marking modified nouns, called Ezafe. It can either mark a given noun or a full noun phrase as a modified element (Samvelian, 2007; Lazard et al., 2006).
Moreover there is an enclitic indefinite article *-i* which doesn't have separate forms for singular and plural; if it attaches to a noun modified by an adjective, it may either directly follow the noun or the adjective. In the first case, the noun does not take the Ezafe particle, whereas in the second case it does (Samvelian, 2007; Lazard et al., 2006). Other enclitic particles are *-i* combined with the relative particle *-ke*, the optional finite marker *-(h)e* and object marking postposition *-râ*. Adjectives only vary in degree by taking the suffixes *-tar* for the comparative and *-tarin* for the superlative form. But they can also take the Ezafe enclitic when following a further modified noun or taking a direct or indirect object. This last point especially holds for adjectives derived from verbal forms (Lazard et al., 2006).
Concerning the verbal class, Persian like most Iranian languages possesses only a very limited amount of verbal lexemes. They form a closed word class of about 200 elements. Most verbal meanings known from the more extensively described Indo-European languages are expressed through complex verbal predicates built from a light verbal head and a predicative element which can be either a noun or an adjective.
Verbal morphology is slightly more complex but follows a rather simple pattern. Persian morphological descriptions usually (Lazard et al., 2006) state the existence of two distinct verbal stems, one for the present tense forms (SI), one for the past tense forms (SII). SI is used in the formation of all present tenses, the present participle, the gerund and the imperative forms whereas SII forms the past tenses, the past participle, the participle of possibility or obligation and the two infinitives. Compound tenses as well as the passive voice are derived from the past participle. All Persian verbal paradigms consist of the combination of a given stem with a set of pre- and suffixes, such as in the following representation:
Modal/Temporal Prefix(es) - Stem - Personal Suffix(es).
The paradigm of temporal/modal prefixes consists of *mi-* and *be-*, respectively for building the indicative and subjunctive/imperative forms. The possible personal suffixes are *-am*, *-i*, *-ad/-e/ø*, *-im*, *-id/-in* and *-and/-an*. These combinations generate seven different tensed forms for six different persons each, as well as five nominal verbforms to which the above mentioned enclitics can be attached. The negational prefixes *n-* or *m-* (more formal, used for the imperative only) can also be attached to the thereby created verbforms (Lazard et al., 2006). Moreover Persian also displays two distinct paradigms of the present indicative forms for the verb *budan* 'to be', one being constitutive of plain words whereas the other is formed of enclitic particles which may attach to nouns or adjectives. This second paradigm is also used attached to the past participle to form both the perfect and imperfect compound tenses (Lazard et al., 2006).
Finally, Persian also has pronominal suffixes which can combine with nouns, pronouns, verbs, prepositions, adjectives and some adverbs (Lazard et al., 2006).

4. The Alexina Framework

We developed the morphological lexicon PerLex within the Alexina framework (Sagot, 2010). This framework covers both the morphological and the syntactic level (e.g., valency), which shall be useful in further stages of PerLex development. Alexina allows for representing lexical information in a complete, efficient and readable way, that is meant to be independent of the language and of any grammatical formalism. It is compatible with the LMF standard[4] (Francopoulo et al., 2006). Numerous resources are already being developed within this framework, such as the Le*fff*, a large-coverage morphological and syntactic lexicon for French (Sagot, 2010), the Le*ff*e for Spanish, SoraLex for Sorani Kurdish, and lexical resources for Galician, Polish, Slovak and soon English.
The Alexina model is based on a two-level representation that separates the description of a lexicon from its use:

- The intensional lexicon factorises the lexical information by associating each lemma with a morphological class (defined in a formalised morphological description) and deep syntactic information; it is used for lexical resource development;
- The extensional lexicon, which is generated automatically by *compiling* the intensional lexicon, associates

[3] In this paper we use a more standard phonetic transcription of the Persian words and affixes that also represents the short vowels and in that sense differs from the described transliteration.

[4] Lexical Markup Framework, the ISO/TC37 standard for NLP lexicons.

each inflected form with a detailed structure that represents all its morphological and syntactic information; it is directly used by NLP tools such as parsers.

In this paper, the syntactic level is left aside, since we are yet building the morphological lexicon.

5. Formalising Persian Morphology

Among the various enclitic particles listed in Section 3.2., not all shall be treated within morphology. In this matter, PerLex makes the following linguistically motivated choices (Samvelian, 2007):

- plural markers, the Ezafe *-(y)e* (when it has a written counterpart), the indefinite marker *-i*, comparative and superlative markers *-tar* and *-tarin*, personal suffixes and enclitic forms of the auxiliary (except when used for building the perfect forms) are considered as inflectional suffixes,
- combined with the relative *ke*, the enclitic particle *-i* forms a compound *_i ke* that is agglutinated to the preceding form,
- other enclitics, including the copula and the *-râ* definite direct object marker, are considered as independent (agglutinated) forms.

Having in mind the linguistic choices mentioned in 3.2., we developed a complete description of Persian morphology in the Alexina morphological language (Sagot, 2007), based on the data in (Lazard et al., 2006). In this format, inflection is modelled as the affixation of a prefix and a suffix around a stem, while *sandhi* phenomena may occur at morpheme boundaries, sometimes conditioned by stem properties. Our description contains in particular 27 verb tables and 5 nouns tables.

6. The Construction of the PerLex Lexicon

Lexical entries have been obtained through the three following steps that we describe below: (1) constructing a basic set of lexical entries using a certain amount of resources; (2) cleaning the obtained lexical entries; (3) adding manually missing entries that were present in the BijanKhan corpus but not yet covered by the already built lexical entries.

6.1. Gathering Lexical Information

We gathered lexical information from various sources, the importance of which is variable among categories:

- the BijanKhan corpus (BijanKhan, 2004; Amiri et al., 2007), an automatically POS-annotated corpus,
- the Persian Wikipedia[5]
- a lexicon of Persian nouns under development by M. Ghassemi at Université Paris-Est (Ghassemi, p.c.),
- the reference grammar of (Lazard et al., 2006),
- introspection by linguists who are native speakers of Persian.

[5]The Persian Wikipedia is available under the following URL `http://fa.wikipedia.org`. We used the extracted *wiki* version of February 16th 2010.

6.2. Building a Base Lexicon

Lexical entries were obtained as follows. A list of verbal lemmas (infinitive form) was developed manually from freely available Internet resources and checked manually. We associated each of them with one (sometimes several) inflection class(es) from our morphological description.

We used the Persian Wikipedia for collecting proper nouns. Those were found through the titles of Wikipedia articles indicating either a city or a person *category*. We collected and normalised the titles of these articles as well as those of all the articles redirecting towards them. We were thereby able to build a lexicon for proper nouns consisting in person and city names which we completed with a list of country names found in the corresponding Wikipedia article. These tasks resulted in a set of over 10,000 proper noun lemmas which all received a inflectional noun class that doesn't allow for the formation of plural forms.

A list of nominal lemmas was extracted from Ghassemi's data. Some of them were already associated with their plural form(s). For others, a statistical look-up in the BijanKhan corpus allowed us to assign their corresponding plural form(s). The result is an inflection class associated with each nominal lemma.

Entries for other categories were extracted from the BijanKhan corpus,[6] and manually completed on the basis of (Lazard et al., 2006). Apart from adjectives, that all receive the unique adjectival class, prepositions and some adverbs (see 3.2.), all entries for these categories are considered invariable, except for some adverbs that may receive the indefinite marker *-i*, the Ezafe or personal suffixes.[7]

6.3. Cleaning the Base Lexicon

At this stage of our lexicon development, and despite the filtering mentioned in footnote 6, there still remains a significant number of incorrect entries, notably among those extracted from the BijanKhan corpus. This follows from the fact that this corpus has been annotated automatically, that is with a nonzero error rate. We therefore removed a certain amount of lexical entries, especially all the nominal and adjectival entries that had been extracted from the BijanKhan corpus and looking like typical plural forms (that is that seemed to end in the plural suffix *-hâ*, followed or not by the Ezafe, the indefinite suffix or the personal suffixes). We also eliminated numerous entries which seemed to contain typographic errors, especially those whose ZWNJ had been omitted or replaced by a white space (such as the incorrect entries for the pronouns *ân hâ* and *ânhâ* instead of correct *ân_hâ*). Finally we suppressed all superfluous characters, such as the diacritic signs for marking vocalisation. The resulting base lexicon forms the first version of our morphological lexicon for the Persian language.

[6]For lemmas with an ambiguous category in the corpus, we discarded entries corresponding to categories that were assigned with a frequency below 1% among occurrences of this lemma. This reduces the amount of noise that comes from annotation errors in the corpus.

[7]Adverbial entries that also exist as adjectives with the same meaning were discarded.

6.4. Expanding the Lexicon

Once this first version of PerLex had been completed, we have searched the BijanKhan corpus for attested word forms that had not yet been taken into account by our lexicon. Those forms may as well be unknown missing forms as they may be incorrect entries or spelling or typographic errors. We therefore sorted them according to their frequency of apparition and manually completed the lexicon from the resulting list. The missing entries were mostly, on the one hand, proper nouns belonging to categories other than those extracted from the Persian Wikipedia (continent names, names of different regions of Iran, etc.), and, on the other hand, several broken plurals. However we were able to observe that a very large amount of unknown forms were in fact due to typographical errors that had not been spotted during the cleaning step.

7. The PerLex Lexicon

The resulting lexicon contains 35,914 lemma-level entries that generate 524,700 form-level entries corresponding to 494,488 distinct forms. Some insight into the distribution according the category is given in Table 1.

Category	intentional entries	distinct lemmas	extensional entries
verbs	171	139	19,776
common nouns	9,553	9,106	177,988
proper nous	10,996	10,938	33,076
adjectives	11,872	11,835	290,537
others	3,322	3,120	3,323
total	*35,914*	*33,454*	*524,700*

Table 1: Quantitative data about PerLex

8. Conclusion and Future Work

We introduced the first version of a large-coverage lexicon for the Persian language. It is now restricted to the morphological level which is currently undergoing manual validation work. The next step of our lexical development work will be to add syntactic information including subcategorisation frames, starting with verbs. In parallel, PerLex will be extended for describing the important phenomenon of complex verbal predicates in Persian. Both these tasks shall be achieved by semi-automatic techniques already used for the development of other Alexina lexicons, and followed by a manual validation step.

PerLex is freely available under an LGPL-LR license on the web page of the Alexina project.[8]

Acknowledgements

This work was supported in part by the PerGram French-German project, funded by the Deutsche Forschungsgemeinschaft and the Agence Nationale de la Recherche (Grant Number MU 2822/3-1) and jointly headed by Pollet Samvelian (Université Paris 3) and Stefan Müller (Freie Universität Berlin).

[8] `http://alexina.gforge.inria.fr`

9. References

H. Amiri, H. Hojjat, and F. Oroumchian. 2007. Investigation on a feasible corpus for Persian POS tagging. In *Proceedings of CSICC'07*, Teheran, Iran.

Jan W. Amtrup, 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.

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

Jon Dehdari and Deryle Lonsdale. 2008. A link grammar parser for Persian. In Simin Karimi, Vida Samiian, and Don Stilo, editors, *Aspects of Iranian Linguistics*, volume 1. Cambridge Scholars Press.

Heshaam Feili and Gholamreza Ghassem-Sani. 2004. An application of lexicalized grammars in English-Persian translation. In *Proceedings of ECAI'04*, Valencia, Spain.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of LREC'06*, Genoa, Italy.

M. M. Hafezi. 2004. A syntactic parser of Persian sentences. In *Proceedings of the 1st Workshop of the Persian Language and Computer*, Teheran, Iran.

Gilbert Lazard, Yann Richard, Rokhsareh Hechmati, and Pollet Samvelian. 2006. *Grammaire du persan contemporain*. Institut Français de Recherche en Iran & Farhang Moaser Edition, Teheran, Iran.

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

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

Behrang QasemiZadeh and Saeed Rahimi. 2006. Persian in multext-east framework. In *FinTAL*, pages 541–551.

Chakaveh Saedi, Yasaman Motazadi, and Mehrnoush Shamsfard. 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.

Benoît Sagot. 2007. Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In *Proceedings of LTC'05*, pages 423–427, Poznań, Poland.

Benoît Sagot. 2010. The L*efff*, a freely available and large-coverage morphological and syntactic lexicon for french. In *Proceedings of LREC'10*, Valetta, Malta.

Pollet Samvelian. 2007. A phrasal affix analysis of the persian ezafe. *Journal of Linguistics*, 43(3):605–645.

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

Samira Tasharofi, Fahimeh Raja, Farhad Oroumchian, and Masoud Rahgozar. 2007. Evaluation of statistical part of speech tagging of Persian text. In *Proceedings of ISSPA'07*, Sharjah, U.A.R.