Developing PersiaNet: The Persian Wordnet

Farhad Keyvan
XselData Corporation, Bridgewater, NJ
fkeyvan@optonline.net

Habib Borjian
Independent Consultant
habibborjian@hotmail.com

Manuchehr Kasheff
Columbia University
mk12@columbia.edu

Christiane Fellbaum
Princeton University
fellbaum@princeton.edu

Abstract

This paper outlines work on PersiaNet, a wordnet for Modern Persian.

Introduction

PersiaNet, like the currently existing WordNets, will constitute a powerful tool for a variety of NLP applications. It will be based on the proven successful design of the Princeton WordNet and will be directly mappable onto the Princeton WordNet as well as the European and Balkan WordNets, allowing for significant crosslinguistic NLP capabilities. Many useful other applications for creating Persian language education materials, dictionaries of Persian mapping to other WordNet languages, and a tool for converting digital Persian text in the current Perso-Arabic alphabet into Latinized Persian transcription, can be created on the basis of Persian WordNet.

Background: The Persian Language

Persian is the major member of the Iranian branch of the Indo-Iranian family of the Indo-European languages. Three major phases are distinguished in its development, namely, Old, Middle and New Persian. Old Persian is represented in the inscriptions of the Achaemenid kings, dating from the 6th century B.C.E., and there is a sizable corpus of works written in Middle Persian. New Persian, now the official language of the three sovereign states of Iran, Afghanistan (where it is referred to as Dari, i.e., formal Persian) and the Republic of Tajikistan, has for centuries served as the main vehicle of culture, literature and politics in a vast area extending well beyond the limits of the Iranian plateau, from the cities of Samarqand and Bukhara in modern Uzbekistan and the Indian Subcontinent in the east to the Caucasus and Anatolia in modern Turkey. The vast corpus of literature produced in New Persian extends over a period of more than a thousand years, representing works that are considered among the masterpieces of the world literary heritage. To mention just a few among many great names, one may recall those of Ferdowsi, the author of the Persian epic, Shah Nameh, Sa'di, Hafez, Khayyam, and Rumi. The language used in the works of the great masters is usually referred to as classical Persian (Dari), which has always enjoyed great prestige and has served as the exemplary and thus the official form of the language. Until a few decades ago, it was the only language that was used in textbooks and studied seriously, although, like any other living language, it differs slightly but noticeably from its colloquial variations.

The use of colloquial language in serious literature began around the turn of the last century and received a forceful social impetus with the advent of the Constitutional Revolution in 1906. Although classical Persian still holds its status of high prestige and is still being studied at every level of education, the literary style of the last few decades has moved closer than ever to the colloquial idiom and has been instrumental in making it gain currency.

Foreign words are found in almost every aspect of the language. The largest group of such terms belongs to Arabic, which was the sole language of learning in the 8th and 9th centuries following the Islamic conquest of Iran, when it functioned like Latin in Europe. Words of Turkish and Mongolian origin are also current in everyday use, the residue of the Turkish and Mongolian dynasties. Moreover, a sizable number of foreign words have entered the language during the past century, particularly in the areas of science and technology: French in Iran, English in Afghanistan, and Russian in Tajikistan.

The cultural and political developments of the last few decades in these three countries are common knowledge. They undoubtedly make the study and the correct understanding of the language and the complex culture it represents, more urgently meaningful than ever before. The creation of a Persian PersianNet is a necessary step towards this end.

The Writing System

Persian has been written in a modified Arabic alphabet since the 9th century AD, when New Persian was first introduced. Ever since, Persian orthography has retained its basic characteristics despite certain modifications during its long history.

Persian uses all Arabic letters plus four consonantal letters not occurring in Arabic. The Arabic pharyngeal and certain dental consonants are not phonemically distinct in Persian, but are retained in all Arabic loans. Hence, there is a choice of two or more letters for certain sounds, e.g., the phoneme /t/ is presented by two distinct letters, /s/ by three, and /z/ by four. The long vowels /i/ and /u/ are represented by the letter of the consonant nearest in pronunciation; thus, the letter *y* represents both /y/ and /i/, and *v* both /v/ and

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 315–318.

/u/. Short vowels may be, but are usually not, represented by diacritics. The main innovation in Persian is that short vowels are always represented by consonant letters in final position: final /o/ by *v*, and final /e/ by *h*.

In texts where optional diacritics are employed, virtually every phoneme of the language is unambiguously represented. In practice, however, the diacritics are omitted, which can make the language difficult for the second language learner. But the native speakers are quite comfortable with the written form of Persian in learning, reading and writing. The process of handwriting in Persian orthography is considered amazingly quick by the Tajiks of the former Soviet Union, who write their form of Persian in the Cyrillic alphabet. (Note that calligraphy has remained a major form of art and has even flourished in the last few decades.) Nevertheless, Persian orthography has been the subject of endless pseudo-scientific arguments in the last 150 years. Romanisation of Persian has long been an ambition of intellectuals, most of whom have lacked an understanding of the linguistic issues involved.

PersiaNet will use Persian orthography as the basic form for both searching and compiling words. It will also employ a parallel Roman writing system as a technography to make it easier to type the Persian word being looked up. Roman letters are necessary because many Internet users, including native speakers of Persian, either are not familiar with the Persian keyboard or find it too difficult to master. Thus our Romanisation is free of diacritic marks and is based on the transcription of the contemporary Persian as used in Iran. Note that the Romanisation alone would result in two major problems: 1. Persian variants of Iran and Afghanistan (and ultimately that of Tajikistan) are phonologically different, particularly in their vowel systems (the differences become unnoticed in the unifying Persian orthography). 2. Persian in Roman symbols yields a large number of homonyms, e.g. the pronounced form /arz/ corresponds to four written forms, each carrying at least one meaning.

Lexicographer's interface

As a first step towards the construction of Persian WordNet, we have developed a web-based lexicographer's interface. The interface is web-based in order to be useful to lexicographer members of the PersiaNet team located in several countries around the globe. We anticipate collaboration with colleagues in Iran.

The PersiaNet project utilizes a single common database for the Persian WordNet that can be accessed and shared by every PersiaNet team member, thus eliminating dependence on disparate databases, which require constant synchronization. Every user will therefore have access to the most up-to-date version of the data. To generate a single Persian WordNet with a standard format and structure. At this initial phase, the effort will be limited to Persian nouns only. The interface will be expanded in its utility to include verbs and adjectives in the future. The interface uses the English WordNet 2.0 ID numbers in connection with corresponding Persian WordNet synsets. Additionally, it should also allow for assigning completely new synset ID numbers. Updates to WordNet version 2.1 will be performed soon. Lexicographers will be able to enter data in both the current Persian script as well as a standard Roman transcription. The interface will be a simple user-friendly GUI which would not require a steep learning curve. The synsets created using the PersiaNet interface will be saved in the XML format for display with disparate software tools available for other WordNet languages. The XML files as well as the data created have to be in Unicode encoding to correctly display the Persian script.

As is apparent from the Figure 1, the GUI consists of two frames. The left frame contains the form with field list and textboxes that are split down the center. This form is used to allow data entry in both Persian script and Roman transcription. The right frame is the synset structure display area used for previewing the complete structure of a synset data entered in the left frame, with all its related antonyms, hypernyms, holonyms, meronyms, and hyponyms. At this point the interface's front-end, plus the back-end database have been created and tested. PersiaNet lexicographers have begun using the interface to generate synsets and populate the Persian WordNet database. The focus is now on increasing the number of lexicographers that could join the project and contribute to the database, taking advantage of the extremely easy-to-use interface.

Lexical Coverage

Synsets

Currently, work on the project is strictly on a volunteer basis. As a result, lexical coverage is very small. The hope is to secure funding by concentrating initially on synsets in areas that are of interest to potential funders. A good strategy for the future would be the one adopted in BalkaNet (Tufis 2004), implementing BalkaNet Concept Sets 1, 2 and 3 (BCS1, BCS2 and BCS3). BCS1 is essentially the Base Concept Set of EuroWordNet (Vossen, 1998), while BCS2 and BCS3 are concepts (Princeton WordNet 2.0 synsets) selected on the basis of their frequency in the six languages of the BalkaNet project. BCS1, BCS2 and BCS3 are conceptually dense, in the sense that for any concept in the BalkanNet Concept Sets all of its hyperonyms (up to the top of the hierarchies) are also in the BalkanNet Concept Sets. This approach has been recently adopted by the development team of the Danish wordnet. (We thank a reviewer for helpful comments on this aspect.)

Domains

Like many existing Wordnets, Persian WordNet includes domains. A domain is lexicalized by a synset, and is linked via a "domain" label to all the synsets that can be said to fall into this domain. We initially extracted the domain labels from the Princeton WordNet. However, these domains do not constitute an optimal set. They are on very different levels of granularity (commerce, exchange) and many arguably important domains are missing (education, chemistry). We propose to construct a domain ontology that covers most of

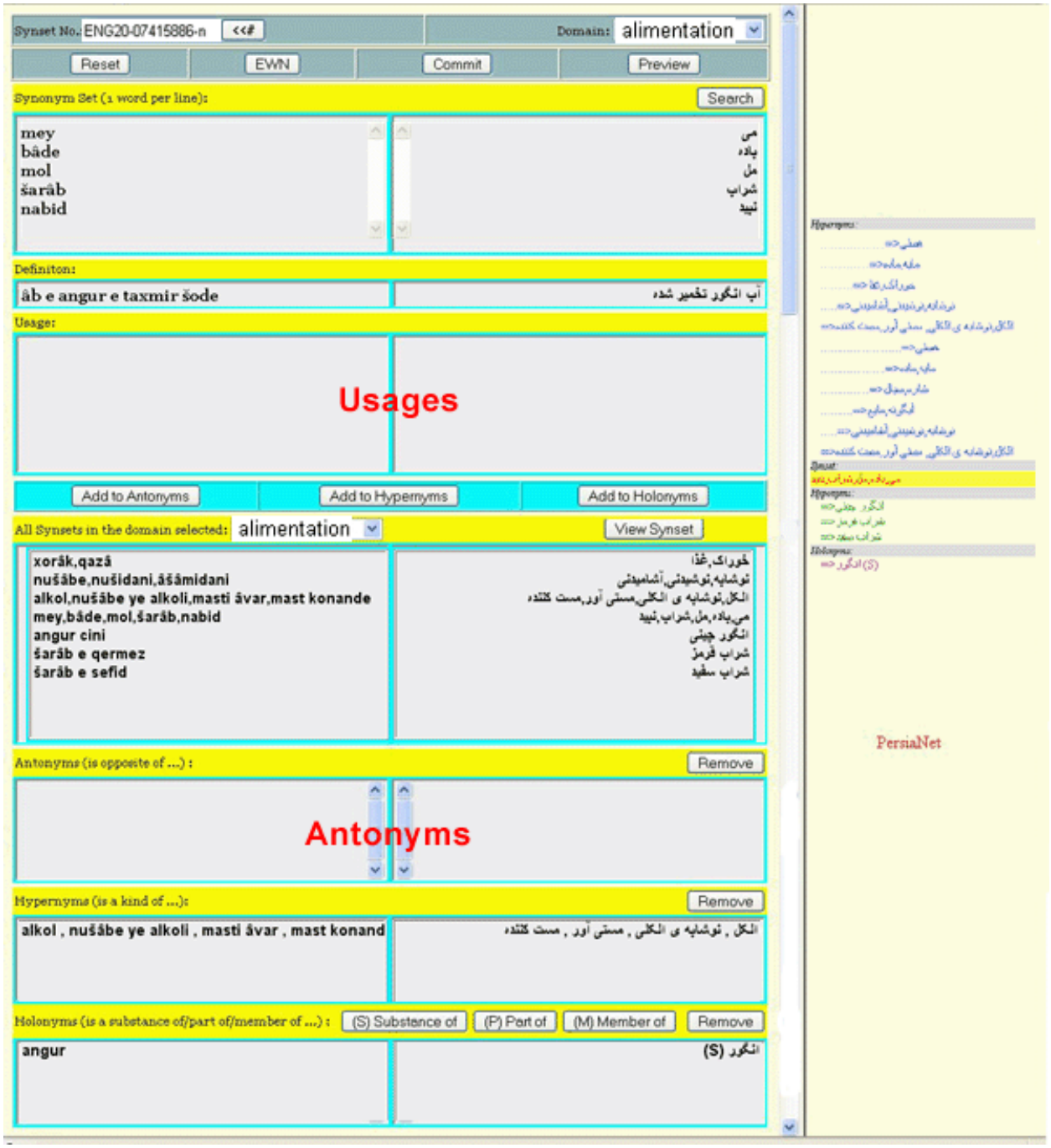

Figure 1: The Lexicographer's Interface

the lexicon and makes the kind of "horizontal" connections among the synsets that cannot be captured by the other, "vertical" relations like hyponymy. Persian WordNet might take advantage of the domain ontology developed for MultiWordNet (Magnini, 2002).

Some domains that have been worked on: sports, transportation, and geography. Some specific problems with mapping to English have been noticed, especially lexical mismatches (one-to-many mappings from Persian to English and vice versa).

Conclusion and Outlook

The work described here sets the stage for a larger-scale lexicographic project whose goal is a good-size database with coverage useful for a broad spectrum of Natural Language Applications, including crosslinguistic information retrieval,

question-answer systems, and machine translation. Because our work has been carried out in an informal setting and entirely on a volunteer basis, lexical coverage is currently very sparse. We are trying to secure funds that will allow us to scale up Persian to a level comparable to other Wordnets. The groundwork addressing the Arabic script (transcription, a Latinizer) will facilitate the construction of Wordnets in other languages using the script, such as the Arabic and Urdu.

References

Borjian, Habib. 1999. *Orthography of Iranian Languages* (in Persian). Tehran.

Fellbaum, Christiane, ed. 1998. *WordNet.* Cambridge, MA: MIT Press.

Geiger, Wilhelm, and Ernst Kuhn, eds. 1895–1904. *Grundriss der iranischen Philologie*, 2 vols. Strassburg: K.J. Trubner.

Magnini, Bernardo and G. Cavagli. 2000. *Integrating subject codes into WordNet.* In: Proceedings of LREC 2, Athens, Greece: ELRA.

Paper, Herbert H., and Mohammed Ali Jazayery. 1955. *The Writing System of Modern Persian*. Washington, D.C.

Schmidt, Ruediger, ed. 1989. *Compendium Linguarum Iranicarum*. Wiesbaden: Reichert Verlag.

Tufis, Dan, ed. 2004. *Special Issue on the BalkaNet Project.* Romanian Journal of Information Science and Technology, Vol. 7, Nos 1–2.

Vossen, Piek, ed. 1998. *EuroWordNet: A Multi-Lingual Database with Lexical Semantic Networks.* Holland: Kluwer.