

ЛЕКСИЧЕСКАЯ ОНТОЛОГИЯ ПЕРСИДСКИХ ГЛАГОЛОВ

А.В. Луканин

Любой язык является сложной системой взаимосвязанных элементов, отношения между которыми можно условно разделить на синтагматические и парадигматические. Данные отношения выявляются на всех уровнях языка и могут быть тем или иным способом формализованы. Все элементы языка вместе с их отношениями представляют собой так называемые лингвистические сети, компьютерная обработка которых даёт возможность решать задачи автоматической переработки письменной и устной речи. Построенная нами лингвистическая сеть фреймов, реализующая синтагматические отношения морфем персидских глаголов и парадигматические отношения между фреймами словоформ, вместе с алгоритмами её обработки, используется для генерации словоформ персидских глаголов из инфинитивов [Lukanin et al. 2007].

Первоначально лексикон созданной нами программы Persian Verb Conjugator (PVC, <http://sartre2.byu.edu/persian/pvc/>), составлял лишь исключения из правил образования формы настоящего времени из формы прошедшего времени. В дальнейшем было решено использовать его в качестве словарной базы так, чтобы пользователи программы могли видеть не только парадигмы запрашиваемого глагола, но и переводы его значений на английский язык. Преобразование ресурса в двуязычный словарь потребовало также указания контекстов употребления глаголов и введения набора отношений между глаголами, учитывающих стили и регистры, семантику и синтаксис персидского языка.

Особый интерес представляет разговорный стиль персидского языка, находящий своё выражение в письменной форме в некоторых художественных произведениях, комиксах, блогах, чатах. В связи с тем, что произношение глаголов в разговорном стиле затрагивает не только основы, но и окончания словоформ, нами были изменены сеть фреймов и алгоритмы её обработки, учитывающие эти различия. Таким образом, программа PVC может сгенерировать парадигмы глагола как в письменном, так и разговорном стиле. Более того, можно выделить несколько разговорных регистров, в которых произношение реализуется по-разному. К примеру, форма глагола «я встал» в письменном стиле – باز ایستادم (bāz istādam), а в разговорном стиле есть 4 уровня реализации (от более высокого к более низкому регистру языка): وایستادم (vāyistādam) – وایستادم (vāstādam) – وایسادم (vāyssādam) – وایسادم (vāssādam). В базе данных PVC соответственно приведены основы для каждого регистра этого глагола и каждый глагол более низкого регистра связан подчинительной связью с глаголом более высокого регистра, так что пользователь может просмотреть парадигмы для любого регистра языка. При этом основы данного глагола в различных регистрах разговорного стиля в словарной базе совпадают, но видоизменяются с помощью правил для генерации словоформ в зависимости от

уровня регистра. В настоящее время программа PVC содержит лексическую базу данных около 2500 глаголов персидского языка, 1300 из которых – глаголы письменного стиля и 1200 – разговорного.

Второй тип отношений, который потребовалось ввести – это гипонимическое отношение между простым глаголом (в данном случае гиперонимом) и составным глаголом (соответственно, гипонимом). Большинство глаголов персидского языка являются составными, образованными из неглагольной части и простого глагола (в базе данных PVC содержится 267 простых глаголов письменного стиля и 123 простых глагола разговорного стиля). Таким образом, имеется возможность просмотреть все согипонимы (составные глаголы) одного гиперонима (простого глагола), что и сделано на отдельной странице сайта: <http://sartre2.byu.edu/persian/pvc/compounds.php>.

Однако образование составных глаголов из неглагольной части и простых глаголов не является случайным. Существуют различные классификации составных глаголов, из которых мы выбрали две: классификацию Энн Лэмбтон и Мохаммада Дабир-Мохаддама.

Энн Лэмбтон [Lambton 1953] берёт за основу своей классификации неглагольную часть, в качестве которой могут выступать:

1. слова персидского происхождения:
 - a. существительное;
 - b. прилагательное;
 - c. предлог или наречие;
 - d. предложная конструкция;
2. слова арабского происхождения:
 - a. существительное;
 - b. причастие;
 - c. прилагательное;
 - d. существительное в сочетании с персидским предлогом.

М. Дабир-Мохаддам [Dabir-Moghaddam 1997] берёт во внимание также и тип простого глагола в составе составного глагола. Он выделяет следующие типы образования:

1. прилагательное + вспомогательный глагол (budan, shodan, kardan);
2. существительное + глагол (kardan, zadan, dādan, gereftan, keshidan, dāshtan, khordan);
3. предложная конструкция + глагол;
4. наречие + глагол;
5. причастие прошедшего времени + пассивный вспомогательный глагол.

Составным глаголам приписаны категории этих двух классификаций, и на странице составных глаголов сайта все глаголы можно группировать по выбранной классификации.

Словарная база является тематической, т.е. лексикограф (пополнение и администрирование словарной базы PVC осуществляет Констанс Боброфф) создаёт необходимый набор рубрик, к которым он относит те или иные

глаголы. Пользователи сайта PVC, соответственно, могут ограничить список показываемых глаголов определённой тематикой (art, banking, time, travel и т.д.). Пользователям сайта PVC предоставлена возможность участвовать в пополнении ресурса, в частности, они могут добавлять примеры употребления глаголов. Каждое добавленное предложение может быть также отнесено администратором сайта к той или иной тематической рубрике, а также к отдельной грамматической рубрике (например, различные варианты употребления показателя дополнения), что позволяет изучающим персидский язык просматривать все предложения, в которых приводится тот или иной аспект грамматики.

Имея достаточно большую базу глаголов, очевидным было указывать не только переводы на английский язык, но и давать ссылки на глаголы-синонимы, что и делалось при пополнении словарной базы. Однако первоначально, как и во многих словарях, синонимы указывали на целые словарные статьи, что создавало трудности при его использовании, т.к. не все значения глагола могут быть взаимозаменяемы в предложениях значениями указанных глаголов-синонимов. Более того, возникли трудности в пополнении словарной базы, т.к. отношение синонимии было установлено между словарными статьями напрямую. То есть, если имелось 5 синонимичных глаголов, требовалось связать каждый глагол с каждым из 4-х его синонимов, а значит, требовалось установить 10 связей (1+2+3+4). Если же к этой группе глаголов добавлялся новый синоним, то требовалось связать его с каждым из 5 его синонимов.

Чтобы решить возникшие проблемы в качестве основной единицы, объединяющей синонимы, нами был взят синсет [Miller 1995], составляющими которого являются синонимичные значения глаголов. Для удобства лексикографа, пополняющего словарную базу программы PVC нами был разработан пользовательский интерфейс (см. рис. 1), где каждому глаголу можно добавить несколько значений, даваемых в виде краткого английского эквивалента, и каждое значение можно включить в один или несколько синсетов. Каждое синонимичное значение сделано в виде прямоугольного блока, в который вписан инфинитив персидского глагола и указано его значение на английском языке. Значение текущего глагола отмечается тёмным фоном. Блоки можно свободно перемещать в рамках одного синсета, предоставляя лексикографу удобство упорядочивания значений глаголов внутри синсетов.

При выборе синсета для значения глагола (рис. 1, ссылка «select a synset» напротив значения глагола) открывается новое окно, где лексикограф может просмотреть все синсеты, создать новый, если необходимо, или найти синсеты, введя инфинитив персидского глагола или его транскрипцию.

Мы планируем также ввести и другие типы отношений (гипонимию, тропонию и т.д.), связывающие между собой синсеты, что приближает нас к созданию WordNet-подобной лексической онтологии. Различными коллективами уже ведутся работы по созданию WordNet для персидского языка [Keyvan et al. 2006; Rouhizadeh 2010], однако наш ресурс обладает некоторыми

преимуществами: результаты работы сразу появляются на сайте программы PVC, которая фактически совмещает в себе несколько типов ресурсов. Это генератор парадигм любых персидских глаголов (в том числе и отсутствующих в нашей словарной базе) и персидско-английский словарь-тезаурус. Кроме того, созданные словарная база и алгоритмы генерации используются также для создания компьютерных тестов по персидскому языку.

Transcription							
Infinitive	Present Stem						
<input type="text" value="sir-āb kardan"/>	<input type="text" value="sir-āb kon"/>						
Note: aa=ā							
Meaning	Tags	add meaning					
<input type="text" value="to quench"/>	<input type="text" value="n/a"/> <input type="button" value="v"/>	select a synset					
<i>drinking</i> [remove from synset]							
<table border="0"> <tr> <td><input type="text" value="سر کشیدن to gulp down"/></td> <td><input type="text" value="سیراب شدن to drink one's fill"/></td> </tr> <tr> <td><input type="text" value="سیراب کردن to quench"/></td> <td><input type="text" value="آب خوردن to drink"/></td> <td><input type="text" value="خوردن to drink"/></td> </tr> </table>			<input type="text" value="سر کشیدن to gulp down"/>	<input type="text" value="سیراب شدن to drink one's fill"/>	<input type="text" value="سیراب کردن to quench"/>	<input type="text" value="آب خوردن to drink"/>	<input type="text" value="خوردن to drink"/>
<input type="text" value="سر کشیدن to gulp down"/>	<input type="text" value="سیراب شدن to drink one's fill"/>						
<input type="text" value="سیراب کردن to quench"/>	<input type="text" value="آب خوردن to drink"/>	<input type="text" value="خوردن to drink"/>					
<input type="text" value="to fully water"/>	<input type="text" value="n/a"/> <input type="button" value="v"/>	select a synset					
<i>irrigation</i> [remove from synset]							
<table border="0"> <tr> <td><input type="text" value="سیراب شدن to be fully watered"/></td> <td><input type="text" value="سیراب کردن to fully water"/></td> </tr> </table>			<input type="text" value="سیراب شدن to be fully watered"/>	<input type="text" value="سیراب کردن to fully water"/>			
<input type="text" value="سیراب شدن to be fully watered"/>	<input type="text" value="سیراب کردن to fully water"/>						

Рис. 1. Интерфейс лексикографа по добавлению значений глаголов в синсеты

Список литературы

1. Dabir-Moghaddam, Mohammad. Compound Verbs in Persian. Studies in the Linguistic Science. 1997. 27(2). Pp. 25–59.
2. Keyvan, F., Borjjan, H., Kasheff, M., Fellbaum, C. Developing PersiaNet: The Persian Wordnet. In Proceedings of the 3rd Global WordNet conference, South Korea. 2006. Pp. 315-318
3. Lambton, Ann K. S. Persian Grammar. — Cambridge University Press, 1953. 330 p.
4. Lukanin, A., Bobroff, C. Frame approach to Persian verb generation for educational purposes. In the Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages. — Linguistic Institute, Stanford, California, USA. July 21-22, 2007. Pp. 98-105.
5. Miller, George A. WordNet: A Lexical Database for English. In Communications of the ACM, 1995. 38 (11). Pp. 39-41.

6. Rouhizadeh, M., Yarmohammadi M. A., Shamsfard, M. Developing The Persian WordNet Of Verbs: Issues Of Compound Verbs And Building The Editor. In the Proceedings of The 5th International Conference of the Global WordNet Association (GWC-2010). Mumbai, India, 2010.