Автоматизация разметки примеров употребления в тезаурусе Persian Verb Conjugator

А.В. Луканин, ЮУрГУ, г. Челябинск

Persian Verb Conjugator (PVC, http://persian.nmelrc.org/pvc)

- генератор парадигм персидских глаголов [1]
- персидско-английский тезаурус глаголов
- значения глаголов группируются в синсеты (наборы синонимов), которые в свою очередь связываются 4 типами отношений: гиперонимия, логическое следование, пресуппозиция и каузация [2]

Примеры употребления глаголов

- являются дополнительным способом изучения персидского языка
- пользователи PVC могут предлагать примеры употребления с указанием транскрипции и источника
- примеры употребления отображаются в тезаурусе после их проверки и отнесения к грамматическим и тематическим рубрикам администратором
- список активных пользователей данной функции: http://persian.nmelrc.org/pvc/sentences.php?contributor

Выделение словоформ глаголов в примерах употребления

- В форму редактирования предложений добавлена функция ручной разметки словоформ (для новых предложений).
- В PVC около 1700 предложений, но т.к. одно предложение может содержать несколько разных глаголов или словоформ одного глагола, количество словоформ во всех предложениях оценивается в 5000 позиций.
- Написана программа полуавтоматической разметки предложений.
- Лексикограф должен подтвердить найденные словоформы либо вручную разметить предложения, если автоматическая разметка произведена неверно.

Анализ через синтез

- Функция генерации словоформ в PVC может генерировать до 449 различных словоформ (инфинитив, 20 видо-временных парадигм по 6 форм (3 лица и 2 числа кроме повелительного наклонения, в котором существует только 2 формы), в активном и пассивном залогах, в письменном и разговорном регистрах).
- Для каждого предложения указывается транскрипция, что удваивает количество возможных словоформ, т.к. необходимо выделять словоформы в обоих полях.
- Для поиска выбрана стратегия совпадений максимальной длины (для каждого глагола генерировались все словоформы и упорядочивались по убыванию длины).
- Например, если для глагола *āmadan «приходить»* в предложении была найдена словоформа *āmade budand* (*3 л., мн. ч., преждепрош. вр.*), то словоформы

to breathe (draw air into, and expel out of, the lungs)
Syn: دمیدن، نفس زدن، دم زدن، استنشاق کردن، تنفس کردن

- entailment
  - draw in (air): نفس کشیدن، نفس داخل دادن، در دمیدن، هوا فرو بردن، استنشاق کردن
  - expel air: نفس بیرون دادن، باز دمیدن، دم بر آوردن
- troponymy
  - breathe noisily during one's sleep: خرخر کردن
  - heave or utter a sigh; breathe deeply and heavily: آه کشیدن، دم بر آوردن
  - breathe noisily, as when one is exhausted: نفس‌نفس زدن
  - breathe with difficulty: خس‌خس کردن

āmade bud (*3 л., ед. ч., преждепрош. вр.*) и *āmad* (*3 л.*, ед. ч., простое прош. вр.) не выделяются в том же месте, но могут быть выделены, если встретятся в ещё неразмеченных частях предложения.

- Однако, в некоторых случаях, одно слово может входить сразу в 2 словоформы. Например, в предложении:

بلد بودن یا نبودن، مسأله این است!

***balad budan** yā **nabudan**, mas'ale in ast!*
Знать или не знать, вот в чём вопрос!

в действительности 2 словоформы: *balad budan* и *balad nabudan*, т.е. слово *balad* нужно было бы выделить дважды. Такие ошибочно размеченные предложения корректируются вручную.

Проверка и постредактирование

- Для удобства ручной проверки результатов автоматической разметки словоформ в предложениях нами была разработана программа, написанная на PHP и jQuery.
- Лексикограф может либо подтвердить правильность разметки, нажав на кнопку Accept, либо указать на ошибку в паре глагол-предложение, нажав кнопку Decline. В последнем случае лексикограф обязан отредактировать эту пару (кнопка Edit).
- Для последующей проверки даётся возможность просмотреть оба варианта разметки. Нажав на кнопку View, можно перейти в карточку глагола с уже опубликованным предложением, где верно выделены словоформы данного глагола.

ZWNJ (zero-width non-joiner, U+200C)

- ставится после приставки mi- в формах прошедшего длительного, настояще-будущего и настоящего определённого времён для разрывного написания символов (например, **می‌دم** (*«давать», 1 л., ед. ч., наст.-буд. вр.*) вместо ненормативного **میدم**).
- Для разметки форм глаголов в ненормативном написании мы добавили в список словоформ для поиска также формы с удалённым символом ZWNJ, что позволило автоматически разметить ещё 133 пары.

Результаты

- Автоматически размечено 2734 из 4002 (68,3%) пар глагол-предложение и выделено 2756 персидских словоформ и 2790 словоформ в транскрипции.
- Сложность выделения словоформ в предложениях связана с тем, что предложения, взятые для иллюстрации из различных источников (блогов, новостных статей) могут содержать написание слов в различных стилях и регистрах: слэнговое, устаревшее или допустимое, но отличающееся от нормы.
- С. Bobroff выверила 2409 пар глагол-предложение, в 2268 парах (94,1%) разметка была произведена верно, 141 пара (5,9%) содержала ошибки.

Библиографический список

1. Lukanin, A. Frame approach to Persian verb generation for educational purposes / A. Lukanin, C. Bobroff // Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages. — Linguistic Institute, Stanford, California, USA. July 21-22, 2007. — P. 98–105.
2. Lukanin A.V. Synset Relationships in the Lexical Ontology of Persian Verbs // Лингвистика в контексте культуры: материалы V Международной научно-практической конференции (Челябинск 28-30 ноября 2012 г.) / под общ. ред. Е.В. Харченко. - Челябинск: Издательский центр ЮУрГУ, 2012. — С. 131–133.