Морфологический словарь личных имен

Назначение словаря

Словарь может использоваться программными системами для автоматического морфологического анализа текста на русском языке. Словарь содержит информацию исключительно о морфологии личных имен.

Содержание словаря

Словарный вход содержит основу личного имени и соответствующие наборы чередований и окончаний для формирования полной словоизменительной парадигмы.

Дополнительно присутствуют следующие признаки:

- пол лица — носителя личного имени (мужской/женский);
- полное имя — указывается для кратких, уменьшительных и т.п. форм (Федя — Федор).

Структура словаря

Словарь хранится в трех файлах:

- файл основ, stems.txt;
- файл векторов чередований, alternations.txt;
- файл векторов окончаний, endings.txt.

Каждая строка файла основ соответствует одному словарному входу. В каждой строке содержится 5 полей, разделенных пробелами:

- текст основы в верхнем регистре;
- пол лица — носителя личного имени (*м* или *ж*);
- индекс вектора чередований (номер строки в файле векторов чередований, отсчет с нуля);
- индекс вектора окончаний;
- текст полного имени или дефис (последний означает отсутствие информации о полном имени.

Каждая строка файла векторов чередований описывает один вектор и содержит 13 полей, разделенных пробелами:

- первое поле содержит слово-пример, в котором имеет место чередование данного вида;
- остальные 12 полей содержат элементы вектора чередований.

Элементы вектора упорядочены следующим образом. Сначала 6 элементов для единственного числа (в порядке именительный, родительный, дательный, винительный, творительный, предложный падеж), затем аналогично описаны 6 элементов для множественного. Каждый элемент вектора предваряется дефисом.

При словообразовании элементы вектора дописываются к основе непосредственно справа.

Каждая строка файла векторов окончаний описывает один вектор и содержит 14 полей, разделенных пробелами:

- первое поле содержит слово-пример, которое имеет окончание данного вида;
- второе поле содержит классифицирующий код;
- остальные 12 полей содержат элементы вектора окончаний.

Элементы вектор упорядочены аналогично векторам чередований и предваряются дефисом. Конструкция "-*" используется для указания на то, что использование такой формы не предполагается (невозможно). Например, родительный падеж множественного числа от имен Лука, Илья.

Некоторые замечания об использовании словаря морфологическими анализаторами

Словарь не содержит составных имен вида *Жан-Клод*. Если составные части таких имен включены в словарь как самостоятельных входы, то морфологический анализ составного имени может быть выполнен с помощью специальных эвристических процедур.

Некоторые личные имена в словаре записаны с буквой *ё* (например, Акиёси). Необходимо учитывать, что в анализируемых текстах возможна подмена *ё* на *е*. Поэтому при загрузке словаря в морфологический анализатор необходимо создавать дубликаты таких имен с буквой *е*.