АВТОМАТИЧЕСКИЙ МОРФОКЛАССИФИКАТОР РУССКИХ ИМЕННЫХ ГРУПП

Большаков И. А. (bolshakov34@mail.ru)
Независимый исследователь, Москва, Россия

Большакова Е. И. (eibolshakova@gmail.com)
Национальный исследовательский университет
Высшая школа экономики, Москва, Россия

Описывается морфологический классификатор русских именных групп с практически произвольным составом и длиной. Раздельно рассматриваются случаи одиночного существительного, группы не более чем с двумя изменяемыми словами и сверхдлинные группы. Классификатор испытан на тех 115 тысячах групп из 1–6 потенциально склоняемых слов, которые входят в словник большого электронного словаря русского языка КроссЛексика. Реализован также автономный вариант, опирающийся на встроенный. Он позволяет склонять именные группы, не входящие в КроссЛексику.

Ключевые слова: морфологическая классификация, русские именные группы, склонение групп, встроенный и автономный морфоклассификаторы

AN AUTOMATIC MORPHOLOGICAL CLASSIFIER OF NOUN PHRASES IN RUSSIAN

Bolshakov I. A. (bolshakov34@mail.ru),
Independent researcher, Moscow, Russia

Bolshakova E. I. (eibolshakova@gmail.com),
National Research University Higher School of Economics,
Moscow, Russia

A morphological classifier of Russian noun phrases of almost arbitrary composition and length is described. Our study covers all peculiarities of Russian declination: fleeting vowels, multiplicity of declination patterns, animacy / inanimacy, specific additional cases, simultaneous occurrence of declinable nouns, adjectives and/or numerals in a phrase. The formula of Russian noun declination is a mere concatenation: word-form= pseudo-stem + pseudo-ending. The declination class is usually determined by the final 1 to 5 letters of the nominative. However exceptions are numerous and they are collected in many built-in lists. We consider separately isolated nouns, groups with no more than two declinable words, and extra long groups. The classifier is a part of CrossLexica, a large electronic dictionary of Russian, and is tested on 115,000 noun phrases of 1 to 6 words included in the CrossLexica dictionary. Based on the built-in classifier, an autonomous module is also constructed, which additionally declines extra-CrossLexica noun phrases.

Key words: morphological classification, Russian noun phrases, phrase declination, built-in and autonomous morphological classifiers.

1. Введение

Практически любая система автоматической обработки текстов нуждается в морфологической классификации обрабатываемых слов. В русском языке для любых слов нужно знать их часть речи и флективный класс (= морфокласс), а для существительных — еще и род. Методы формального описания русской морфологии давно известны. Хороший обзор по теме содержится в работе [Sokirko], к которой мы и отсылаем.

Когда задумывался морфологический классификатор для большого словаря русских словосочетаний и смысловых связей КроссЛексика [Bolshakov], уже был внедрен в русскоязычную .версию Microsoft Word словарь А. Зализняка [Zaliznjak], а в двуязычных словарях МультиЛекс использовалась полиморфная модель А. Гельбуха [Gelbukh]. Однако задача классификации новых слов, не входящих в словари, еще не была решена. Не существовало и больших

текстовых корпусов и массивов Интернета, для которых такая задача должна была неминуемо возникнуть.

Для КроссЛексики был нужен морфоклассификатор элементов словника, не известного заранее ни по составу, ни по объему. Классификатор должен был осваивать новые морфологически не размеченные слова по мере их ввода в систему. Возникшая классификация предназначена для образования нужной флективной формы и полной морфологической парадигмы слова. При этом допускалось многолетнее совершенствование морфологических средств в процессе эволюции КроссЛексики.

Наша задача была сходна с задачей предсказания морфологии [Sokirko], но имела важнейшую особенность: объектом классификации могло быть как одиночное слово, так и целое словосочетание. В данной работе рассматриваются именные словосочетания (далее — именные группы, ИГ).

ИГ составляют 42 % словника КроссЛексики и включают слова разных частей речи. В любом контексте ИГ играет синтаксическую роль существительного. ИГ может быть как одиночным существительным, так и словосочетанием из существительных, прилагательных, числительных и слов иных частей речи: ***точка*** *зрения,* ***вид*** *на жительство,* ***право*** *быть избранным,* ***свободная экономическая зона****, свободно* ***конвертируемая валюта, семь смертных грехов****,* ***теория*** *вероятностей и* ***математическая статистика***.

Внутри ИГ может быть несколько параллельно склоняемых слов (выше выделены шрифтом). Паттерны склонения для всех этих слов сводятся нами в единую таблицу, т. е. формально все они причисляются к существительным. Неизменяемость слова в ИГ понимается как сохранение вида словоформы при склонении группы в целом. Например, словоформа *речи* в группе *часть речи* является генитивом от *речь*, но эта форма сохраняется при любом вхождении объемлющей ИГ в контекст. Наличие нескольких изменяемых слов требовало синтаксического анализа элементов словника, а мы хотели его избежать. Облегчающим обстоятельством явилось то, что входной формой ИГ является ее именительный падеж.

Итак, на вход нашего морфологического классификатора подается ИГ в номинативе, а на выходе получается набор морфоклассов входящих в нее слов. Вся совокупность результатов классификации формирует внутренний морфословарь КроссЛексики. В рабочем режиме он позволяет восстанавливать любые косвенные падежи ИГ и выводить на экран всю морфопарадигму ИГ из шести падежей. На данный момент внутренний морфословарь включает 115 тыс. ИГ, наиболее употребимых в русском языке. Большинство многословных ИГ взято из Интернета.

Строя собственную морфологическую модель одиночного русского существительного, мы исходили из простейшей конкатенации <словоформа> = <псевдооснова> + <псевдоокончание>. Базовой же идеей предсказания морфокласса был принцип аналогии Г. Белоногова [Belonogov]: одинаковые цепочки последних букв (= финали), как правило, соответствуют одинаковому морфоклассу. Хотя финали длиной до пяти букв использовались широко, пришлось создать около сотни встроенных списков разной длины из слов,

противоречащих аналогии. Для минимизации общего объема списков привлекалась статистика существительных разных морфоклассов.

В результате многолетней эволюции КроссЛексики, порядка 15% одиночных существительных ее словника оказались отсутствующими в словаре Зализняка, но в рамках нашей морфосистемы склоняющимися правильно. Таким образом, нами построена морфосистема, альтернативная и эквивалентная системе словаря [Zaliznjak] (если не учитывать ударений и буквы «ё»). При этом наша система опирается на существенно обновившуюся за последние десятилетия лексику русского языка.

Из внутреннего морфословаря КроссЛексики и обслуживающих его подпрограмм был дополнительно сформирован автономный морфоклассификатор. Пользователь вручную вводит в него произвольную ИГ, а на экране получает ее правильную морфопарадигму. Кроме ИГ, содержащихся в КроссЛексике, автономный классификатор способен правдоподобно обрабатывать именные группы, отсутствующие в КроссЛексике, а также буквенные цепочки, отсутствующие в языке.

2. Классификация одиночных существительных

Простота приведенной выше формулы конкатенации обманчива, и морфоклассификатор русских существительных должен иметь многократно больше морфоклассов, чем указано в академических грамматиках.

- Большое количество существительных имеют беглую гласную. Поскольку беглых гласных в русском языке две, и порой они заменяются на *й* или ь, а следующих за ними разных согласных 13, для правильного морфологического описания нужны несколько десятков дополнительных классов.
- Существительные бывают одушевленные и неодушевленные. Их раздельное описание удваивает число морфоклассов.
- Существительные единственного и множественного числа мы рассматриваем раздельно, каждое со своей морфопарадигмой. Эта мера оправдана не только комбинаторными свойствами слов разных чисел, но и их различиями в наборах семантических связей. Число морфоклассов еще раз удваивается.
- Существительные, склоняющиеся по типу прилагательных, столь многочисленны, что покрывают все три грамматических рода, оба числа и практически все классы склонения прилагательных. Это прибавляет еще 46 классов.
- Именные группы могут формироваться числительными, у младших из которых одушевленный и неодушевленный варианты различны. Это еще десяток классов.
- Специфическим случаям типа *путь, дитя* или *Христос* приходится отводить отдельный класс. В итоге число морфоклассов достигло 230.

Дополнительно введены классы для существительных, входящих в многословные числовые группы типа *десять заповедей*. Здесь существительное *заповедей* имеет весьма необычную совокупность падежных окончаний {*ей, ей, ям, ей, ями, ях*}. Такие паттерны вынудили добавить еще 25 классов. Они эффективно обслуживают без привлечения сложных синтаксических правил множество ИГ, управляемых «малыми» (2–4, 22–24, 32–34...) и «большими» (5–20, 25–30, 35–40...) числительными в виде цифр либо слов.

Приведем фрагмент таблицы окончаний различных морфоклассов:

```
/*1*/{"","а","у","","ом","е","у","у"},   /*порт,бак,барыш*/
/*2*/{"","а","у","а","ом","е","а","е"}, /*автор,критик,малыш,подлец*/
/*3*/{"","а","у","","ем","е","у","у"},   /*абзац,вкладыш*/
/*4*/{"","а","у","а","ем","е","а","е"}, /*муж,гаденыш*/
/*5*/{"","а","у","а","ым","е","а","е"}, /*Иванов,Путин*/
/*6*/{"а","ы","е","у","ой","е","ы","е"},/*лапа,папа*/
/*7*/{"а","ы","е","у","ей","е","ы","е"},/*улица,певица*/
/*8*/{"а","и","е","у","ой","е","ы","е"},/*река,нога,межа,девушка*/
```

Слева идут номера морфоклассов, справа — примеры существительных. Число падежей увеличено до 8, поскольку у небольшого, но достаточно частотного ряда существительных мужского рода есть частичный (*выпить* ***чаю***) и / или местный падеж (*в* ***аэропорту***). В классах, где дополнительных падежей нет, их окончаниями взяты таковые родительного и предложного падежей соответственно. Особый вариант аккузатива в выражениях типа *пойти в программисты* считается нами номинативом с особым предлогом *в4* и с заглавным вопросом «*пойти* ***в какие люди***?» в модели управления.

С аналогиями среди существительных тоже не все гладко. Так в паре {*восток, росток*} слова с одинаковой финалью длины 5 склоняются поразному, но хотя бы имеют одинаковые род и число. Однако, например, в паре {*метель, мотель*} с одинаковой финалью длины 4 род различен. У омонимов нередко бывает разная одушевленность: *оператор1* ‘человек’ Vs. *оператор1* ‘операция’, а в паре *виски1* ‘части лица’ Vs. *виски2* ‘напиток’ различны и род и число.

Соображения аналогии пришлось дополнить оценочной статистикой крупных классов существительных. Упомянем лишь три учтенных нами наблюдения.

- Неодушевленных существительных больше, чем одушевленных в 2–6 раз, в зависимости от финали. Поэтому, если вводится новое слово и одушевленность его известна, мы помещаем его во встраиваемый список соответствующего класса, а по умолчанию классифицируем слово более правдоподобным способом как неодушевленное.
- Существительных множественного числа с финалью *ы*, склоняющихся по мужскому типу, заметно больше, чем склоняющихся по женскому типу. Поэтому, если появляется новое слово с женским типом склонения, оно

помещается в списки раздельно одушевленных и неодушевленных слов, а по умолчанию новое слово на *ы* классифицируется по мужскому типу.
- Существительных женского типа с финалью <нешипящая согласная>+*ки* (*доски, пробки*…) больше, чем подобных мужских (*броски, ростки*…). Поэтому в списки вносятся слова мужского типа, а по умолчанию новому слову присваивается женский.

В результат возникли порядка сотни встроенных списков, в основном, коротких. Наиболее длинный из них содержит одушевленные существительные единственного числа на твердую согласную (2250 слов). Встроенными списками суммарно покрываются примерно 12 тыс. слов, всего же в КроссЛексике сейчас около 48 тыс. одиночных существительных, так что нашим классификатором обеспечивается экономия примерно в четыре раза.

С учетом сказанного разработана процедура **SINGLE**(*noun, cls, num, gen*), где *noun* — входное существительное, например, *стена, стены, няни, нянечки, окно, вектор, строение, строения, учитель, учителя, лектор, студенты, студентки, бычки, шоссе, рагу, Горбачев, Михаил, Гарри, МГУ.* Остальные параметры процедуры выходные: *cls* — морфокласс существительного: 0,…,255*; num* — его число: единственное или множественное; *gen* — род: мужской, женский, средний или общий (только для множ. числа). Хотя в большинстве своем число и род однозначно определяются морфоклассом, это не так, например, для существительных на ***а*** типа *папа* или *бомбила* и для аббревиатур. Формируемые род и число ИГ нужны в разных частях КроссЛексики для проверки согласования.

Процедура **SINGLE** сначала разделяет существительные по их последним буквам. Далее существительные делятся исходя из предпоследней буквы. Далее используются более длинные финали и встроенные списки. Для примера, существительное *мама* получает правильный морфокласс вместе с сотнями других слов с финалью *ма* после предварительной проверки отсутствия в нем финали *дома* (как у слов *дома, автодома, детдома*…) и несовпадения со следующими словами: 1) *полкилограмма*, 2) *бельма*, 3) *письма*, 4) *Дюма*, 5) из короткого списка *далай-лама, Дима, Ерема, Кузьма, Обама*… 5) из короткого списка *закрома, корма, терема, тома*… После всех проверок с отрицательным результатом слову *мама* присваивается нужный класс по умолчанию.

3. Проверка и исправление морфокласса

Классификация одиночных существительных процедурой **SINGLE** не безупречна:
- Омонимам разной одушевленности (*конструктор, оператор, бычки*…) присваивается одинаковый неодушевленный класс, поскольку **SINGLE** не учитывает омонимии.
- Существительные, склоняемые как прилагательные, признаются одушевленными при наличии их во встроенном списке *обрученные, разнорабочие,*

трудящиеся..., а по умолчанию считаются неодушевленными. В итоге, например, слово *военные* будет классифицировано как неодушевленное, что верно для группы *военные расходы*, но неверно при отдельном употреблении слова или в группе *демобилизованные военные*.

- Новые существительные множественного числа на *ы* по умолчанию относятся к мужскому неодушевленному типу (как *столы*), что не всегда правильно.
- Новые существительные с финалью <нешипящая согласная> + *ки* по умолчанию относятся к женскому неодушевленному типу (как *доски*), что тоже не всегда верно.

В то же время КроссЛексика включает обширные ресурсы, позволяющие проверить и даже исправить класс, присвоенный процедурой **SINGLE.** Это обширная база коллокаций (в целом их более 2 млн.), среди них типа *VN* (глагол → существительное: *нанять сотрудника*) и *NN* (существительное → существительное: *наем сотрудника*).

Вот как отражены в исходной текстовой базе КроссЛексики *VN*-коллокации существительного *оператор1*:

```
оператор1/«человек»
  возлагать ... на1 ~а
  заменить ~а
  использовать ~а
  нанять ~а
  обратиться1 к ~у
  передать2 ... ~у
  посадить4 ~а
  пригласить ~а ...
  работать ~ом
  работать с2 ~ом
  стать ~ом
```

Цифры в конце слов означают номера омонимов, а тильда означает (псевдо) основу.

Для исправления неверного морфокласса используются противоречия окончаний существительного, записанных в базе коллокаций, и окончаний, соответствующих морфоклассу, определенному процедурой **SINGLE.** Противоречия возникают по двум причинам: (1) составителем базы коллокаций неверно введено окончание либо (2) **SINGLE** неверно определил класс. Чтобы автоматически различать эти случаи, используется булев показатель ***валидности***, дополнительно вырабатываемый процедурой **SINGLE** для каждого классифицируемого существительного: слово **валидно,** если в его морфоклассе нет сомнения в любых контекстах. Заметная часть валидных слов включено во встроенные списки.

При возникновении противоречий в случае валидного слова диагностируется ошибка составителя базы, и он должен исправить ошибку вручную.

В случае невалидного слова при отсутствии противоречий оно считается **верифицированным**. Если же противоречия есть, то программа сначала пытается исправить назначенный класс сама, без выдачи сообщений составителю базы коллокаций. Наиболее часто при этом неодушевленный класс исправляется на одушевленный.

Например, для слова *оператор1* при первоначально назначенном неодушевленном классе 1 обнаруживается противоречие в коллокации *возлагать на оператора* (здесь было бы тогда правильным *возлагать на оператор*). Противоречивы и коллокации *заменить / использовать / нанять оператора,* но автоматически диагностировать это без предлога невозможно. В КроссЛексике имеется список из 63 соответствий {неодушевленный класс Vs. одушевленный класс}, где классу 1 противопоставлен одушевленный класс 2. Этот список и используется для коррекции: замена класса 1 на класс 2 устраняет для слова *оператор1* как обнаруженные, так и необнаруженные противоречия. Слова, у которых класс исправлен указанным способом, назовем **анимизированными**.

Существуют и другие исправимые противоречия. Так, если не включить существительное *лакуны* во встроенный список, можно исправить его «мужской» класс, полученный по умолчанию, на «женский», обнаружив пустое окончание в форме *лакун* (вместо **лакунов*). Слова, исправленные этим способом, назовем **феминизированными**.

Исправить класс слов *броски* или *ростки* с «женского» на «мужской» можно благодаря наличию для них в базе КроссЛексики псевдоокончаний *ов/ам/ами/ах* вместо *ок/кам/ками/ках*. Так исправленные слова назовем **маскулинизированными**.

Проверка окончаний для коллокаций типа *NN* не обеспечивает автоматических исправлений, но обнаруженные здесь противоречия тоже позволяют исправить базу коллокаций и / или встроенные списки**.**

После отладки встроенных списков и базы КроссЛексики получена следующая статистика валидации и верификации:

Валидные	57,7 %
Верифицированные	40,5 %
Анимизированные	1,6 %
Феминизированные	0,1 %
Маскулинизированные	0,1 %

Три последних позиции минимизированы переносом автоматически исправленных слов в соответствующие списки тогда, когда это допустимо. Но совсем исключить эти позиции принципиально невозможно из-за омонимов разной одушевленности и из-за возможной зависимости морфокласса от контекста.

Остается упомянуть еще омонимы, исправлять классы которых лучше подругому. В первую очередь это несклоняемые слова *ателье, бюро, интервью, казино, кафе, такси, фото, шоссе*... Условно мы приняли их единственное число первым омонимом, а множественное — вторым. **SINGLE**, не учитывающий омонимии, присваивает обоим омонимам единственное число, а затем второму омониму по номеру присваивается множественное число.

В итоге образуется файл BASICS из 20,6 тыс. одиночных классифицированных существительных, наиболее частотных в русском языке. Прочие процедуры КроссЛексики уже не классифицируют его составляющие заново. Пока

не попавшие в BASICS 32 тыс. более редких одиночных существительных КроссЛексики постепенно переходят сюда по мере расширения базы коллокаций, а редкие одиночные пополняются новыми извне.

4. Классификация групп не более чем с двумя изменяемыми словами

Среди разнообразия ИГ в первую очередь выделим те, которые содержат не более двух изменяемых слов. Обнаружено четыре типа таких групп:

Дефисная группа. Разделителем частей ИГ является первый дефис, до которого нет пробела. Левая часть — изменяемое слово, правая часть — единственное слово, изменяемое или неизменяемое (*директор-распорядитель, шкаф-купе, страна-участница*). Сюда не относятся случаи слов с неразрывным дефисом типа *вице-президент*, *топ-менеджер* или *интернет-магазин*. Для различения этих случаев используются эвристики.

Сочиненная группа. Разделителем частей группы является сочинительный союз *и, или, да*. Обе части — одиночные слова, и любое может склоняться (*мама и папа, аукцион или конкурс, совет да любовь, готика и барокко*). Данный вариант покрывает и группы с сочинительным сокращением типа *аудио- и видеопродукция*.

Согласованная группа. Разделителем частей служит пробел. Слева и справа от пробела — одиночные слова, изменяемые или не изменяемые, которые согласованы по роду и числу (*конвертируемая валюта, конструкторское бюро, сочинительный союз, социальное государство, царь Петр*).

Левоуправляемая группа. Разделителем частей служит первый пробел. Левая часть — одиночное изменяемое слово, правая часть состоит из любого числа неизменяемых слов (*часть речи, вид на жительство, пресс-секретарь Белого дома*). Суммарное количество слов в таких группах может произвольно превышать установленный ранее предел 6: *Организация по безопасности и сотрудничеству в Европе.*

Именные группы, входящие в один из указанных типов, многочисленны: в КроссЛексике их 50 тыс. Технически оказалось затруднительным включить в указанные типы, например, группы *здоровый образ жизни, аутсорсинг и управление проектами, государства-участники СНГ, жилой фонд Москвы*, хотя в них ровно два изменяемых слова. Они отнесены к сверхдлинным группам.

Процедура **DOUBLE** служит для классификации ИГ описанных выше типов. Она получает на вход ИГ из произвольного числа слов и выдает один или два морфокласса, а также число и род, характеризующие группу в целом. **DOUBLE** последовательно проверяет соответствие ИГ одному из определенных выше типов. Когда соответствие найдено, к левой, правой или обеим частям группы применяется процедура **SINGLE**, что выявляет выходные морфоклассы, а также род и число группы в целом.

В качестве числа и рода дефисной группы берутся таковые левой части, например, *самолет-амфибия* — ед. муж., *ракета-носитель* — ед. жен.,

общество-паразит — ед. сред. В некоторых сочиненных группах, перечисленных списками, обе части имеют единый денотат, и тогда группе в целом присваивают его род, мужской (*друг и партнер, лжец и болтун*) или женский (*подруга и соперница*), а число берется единственным. Но по умолчанию денотаты различны, и группе присваивается множественное число.

В завершение **DOUBLE** проводит уточнение полученных параметров, бесконтекстное, диктуемое всей совокупностью хранимых в КроссЛексике многословных ИГ, либо контекстное. Так существительные *жучки / клещи / дипломаты / нищие / святые* всегда одушевляются, *банки / риски / белки* переводятся в склоняющиеся по мужскому типу, а *корма* и *вина* переводятся во множественное число. Наиболее частым приемом контекстного уточнения является анимизация левой части согласованной ИГ за счет правой части (*научные работники, лучшие люди*). В дефисных группах правая часть деанимизируется за счет левой (*город-герой*). Для многословных ИГ тоже приходится прибегать к встроенным спискам. Особо велик и пока еще растет список левых частей левоуправляемых групп (4500+ компонентов).

5. Классификация сверхдлинных групп

Среди многословных ИГ КроссЛексики встречаются и те, которые не охвачены описанными выше типами или содержат более двух изменяемых слов. Назовем такие ИГ сверхдлинными. Особо большое их количество оказывается среди десятков тысяч сочиненных пар, заимствованных из Интернета. При их морфологическом описании пришлось ограничить их состав. Конкретно, мы рассматриваем ИГ не более чем из шести слов, любое из которых может склоняться.

В целом они включают:

- ИГ с **тремя** изменяемыми словами (несколько тысяч): *Великая отечественная война, антибиотики и сульфаниламидные препараты, банковская гарантия и поручительство, логическое и творческое мышление*.
- ИГ с **четырьмя** изменяемыми словами (несколько сотен): *Великая октябрьская социалистическая революция, артезианская скважина и водонапорная башня, свободные экономические зоны и льготы*.
- ИГ с **пятью** изменяемыми словами (несколько десятков): *зеленые насаждения и малые архитектурные формы, десять заповедей и семь смертных грехов, единый социальный налог и страховые взносы, активная жизненная позиция и позитивный настрой*.

Статистически значимые примеры со всеми шестью изменяемыми словами типа *первый новый московский опытный моторный завод* в Интернете пока не встретились. В указанных ограничениях коллекция сверхдлинных ИГ в КроссЛексике открыта для пополнений. Однако внесение в словарь новой сверхдлинной ИГ требует ввода вручную в специальный текстовый файл EXLONG отдельной строки, содержащей как саму ИГ, так и некоторые

ее показатели, позволяющие избежать синтаксического анализа. Вот примеры строк этого файла:

```
1 1 1 9 9 9 1 2 Великая отечественная война
1 0 1 1 9 9 2 4 антибиотики и сульфаниламидные препараты
1 1 0 1 9 9 2 4 банковская гарантия и поручительство
1 1 1 1 9 9 1 2 Великая октябрьская социалистическая революция
1 1 0 1 1 9 2 4 артезианская скважина и водонапорная башня
1 1 1 0 1 9 2 4 свободные экономические зоны и льготы
1 1 0 1 1 1 2 4 десять заповедей и семь смертных грехов
1 1 1 0 1 1 2 4 единый социальный налог и страховые взносы
1 0 1-0 9 9 2 4 кровати и шкафы-купе
```

Начальные *n* цифровых позиций, 3 <= *n* <= 6, кодируют изменяемость / неизменяемость (1/0 соответственно) каждого слова ИГ из *n* слов. Седьмая позиция кода занята показателем числа, восьмая — показателем рода ИГ в целом. Так, группа *Великая отечественная война* имеет единственное число и женский род, а большинство групп имеют множественное число и общий род. Заметим, что разделителем слов может быть не только пробел, но и дефис, смотри последнюю строку примеров. На данный момент файл EXLONG содержит 9840 строк.

На основе файла EXLONG процедура **MULTIPLE** обработки сверхдлинных ИГ каждому изменяемому слову в группе назначает его морфокласс, последовательно обращаясь к процедуре **SINGLE**. Назначенные номера классов запоминаются в файле EXLONG вместо показателей изменяемости «1». Далее совершаются автоматические бесконтекстные и контекстные правки, подобные указанным выше. Чаще всего — это анимизация слов, непосредственно предшествующих одушевленным словам.

Чтобы сформировать полную морфопарадигму сверхдлинной ИГ, из файла EXLONG извлекается ее последовательность классов склонения и по ним образуется вся форма со словами в одинаковом падеже. Пример полной морфопарадигмы дан ниже, где изменяемые окончания выделены:

имен *десят**ь** заповед**ей** и сем**ь** смертн**ых** грех**ов***
род *десят**и** заповед**ей** и сем**и** смертн**ых** грех**ов***
дат *десят**и** заповед**ям** и сем**и** смертн**ым** грех**ам***
вин *десят**ь** заповед**ей** и сем**ь** смертн**ых** грех**ов***
твор *десят**ью** заповед**ями** и сем**ью** смертн**ыми** грех**ами***
пред *десят**и** заповед**ях** и сем**и** смертн**ых** грех**ах***

6. Автономный морфоклассификатор

Автономный морфоклассификатор содержит в качестве базы внутренний морфословарь КроссЛексики. Процедура-движок сначала проверяет,

имеется ли в нем входная группа. Если да, то группе присваиваются уже известные параметры.

Если ИГ в морфословаре отсутствует, то ее классификация выполняется процедурами **SINGLE** и **DOUBLE**. Правильно присваиваются классы огромному количеству пока не включенных в КроссЛексику групп произвольной длины. Так, еще до включения в КроссЛексику, автономный морфоклассификатор просклонял правильно малоизвестные пока слова *диспорт* и *угги* и новое понятие *арабская весна*. Группа из 8 слов *перечень наркотических средств, психотропных веществ и их прекурсоров* в КроссЛексике отсутствует, но она тоже склоняется правильно.

Заодно правильно склоняются многие не существующие в языке группы, включая пресловутое словосочетание Щербы *глокая куздра*. Поэтому мы вправе заявить, что обрабатываемые группы имеют практически произвольный состав и длину. В литературе морфоклассификаторы многословных групп пока не известны.

Отметим неидеальную обработку автономным классификатором тех употребленных отдельно омонимов, у которых парадигма зависит от номера омонима. Внутри КроссЛексики правильно их классифицировать помогают именно эти номера. Требовать их знания от пользователя нельзя, и поэтому классификатор присваивает омониму наиболее правдоподобный класс. Полагая, что новые поступления будут удовлетворять приведенной выше статистике валидации и верификации, мы ожидаем правильный класс нового одиночного существительного в 98 % случаев, а маловероятная ошибочная классификация проявится, скорее всего, только в винительном падеже.

7. Заключение

Предложен морфологический классификатор, сформированный в процессе пополнения объемлющей системы — большого электронного словаря словосочетаний КроссЛексика. Классификатор совершенствуется уже более 16 лет и в настоящее время практически стабилизировался в части алгоритма и встроенных списков. Он испытан на всех 115 тыс. именных групп, входящих сейчас в КроссЛексику. При этом для 46 % групп достаточна процедура **SINGLE**, для 45 % нужна **DOUBLE,** а для остальных 9 % — **MULTIPLE**. Алгоритм и списки таковы, что ошибки классификации обнаружить уже не удается.

Автономный морфоклассификатор позволяет классифицировать с ничтожным процентом ошибок как включенные, так и не включенные пока в КроссЛексику именные группы и даже цепочки, отсутствующие в языке, но морфологически его имитирующие.

Наши классификаторы ориентированы на генерацию падежных форм. Анализа косвенных форм ИГ осуществим только совместно с другими частями речи. Пока в КроссЛексике анализируются только любые одиночные слова.

Выражаем признательность А. Ф. Гельбуху за встраивание морфологических процедур в КроссЛексику.

References

1. *Belonogov G. G.* On the use of a similarity method for automatic processing of textual information [Ob ispol'zovanii metoda analogii pri avtomaticheskoj obrabotke tekstovoj informatsii] Problems of cybernetics [Problemy kibernetiki]. Issue 28. Moscow: Nauka Publ., 1974.
2. *Bolshakov I. A.* CrossLexica: A large electronic dictionary of collocations and semantic links between Russian words [KrossLeksika — bolshoj èlektronnyj slovar' sochetanij i smyslovykh svjazei russkikh slov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2009" [Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2009"]. Moscow, 2009, pp. 45–50.
3. *Gelbukh A. F.* Effectively realizable morphologic model of inflective language [Èffectivno realizuemaja model' morfologii flektivnogo jazyka]. Nauchno-tekhnicheskaja Informatsija [Scientific and Technical Information], series 2, #1, 1992, pp. 24–31.
4. *Sokirko A. V.* Bystroslovar': morphological prediction of new Russian words using very large corpora [Bystroslovar': predskazanie morfologii russkikh slov s ispol'zovaniem bolshikh lingvisticheskikh resursov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" [Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2010"]. Moscow, 2010, pp. 450–45
5. *Zaliznjak A. A.* Grammatical dictionary of Russian: Inflection [Grammaticheskij slovar' russkogo jazyka: Slovoizmenenie]. Moscow: Russkij Jazyk Publ., 1977.