
НА'СПЕХ, нареч. Торопливо, поспешно. Делать н. Н. пересуетиться.
НАСПИРТОВА'ТЬСЯ (-тужусь, -туешься, 1 и 2 л. не употр.), -туется; сов. Пропитаться спиртом. | несов. наспиртовывать-ся (-аюсь, -аешься, 1 и 2 л. не употр.), -ается.
НАСПЛЕ'ТНИЧАТЬ, -аю, -аешь; сов. (разг.). Наговорить сплетен о ком-чём-н. Кумушки насплетничали.
НАСТ, -а, м. Заледеневшая корка на снегу после короткой оттепели. Скользить по насту. | прил. на'стовый, -ая, -ое.
НАСТАВИ'ТЕЛЬНЫЙ, -ая, -ое; -лен, -льна. Содержащий наставление (во 2 знач.), поучительный. Н. тон. | сущ. наставительность, -и, ж.
НАСТИЛА'ТЬ, -ю, -ешь; сов. То же, что стлать (во 2 знач.). | прич. настилочный, -ая, -ое.
НАСТИЛО'ЧНЫЙ см. настил, настилать и стлать.
НАСТИ'ЛЬНЫЙ, -ая, -ое; -лен, -льна (спец.). 1. Плавный, идущий почти по прямой линии. Н. прыжок на лошади. 2. Идущий параллельно поверхности земли на определённой высоте. Настильная траектория полёта пуль. | сущ. настильность, -и, ж.
НАСТИРА'ТЬ, -а́ю, -а́ешь; -йранный; сов., что и чего (разг.). Выстирать в каком-н. количестве. Н. целый узел белья. | несов. настирывать, -аю, -аешь.
НАСТИ'ЧЬ и НАСТИ'ГНУТЬ, -йгну, -йгнешь; -йг и -йгнул, -йгла; -йгни; сов. 1. кого-что. Догнать, преследуя. Н. беглецов. Пуля настигла врага (перен.: поразила). 2. (1 и 2 л. не употр.), перен., кого (что). О тягостном событии: внезапно наступить. Беда настигла неожиданно. Альпинистов настиг снегопад. | несов. настига́ть, -а́ю,

СИНТЕЗ ТЕКСТА

СЕРГЕЙ СЛЕПОВ
sergey@morpher.ru

Речь существует в двух основных формах: устной и письменной. Синтезу устной речи посвящено много исследований и статей, а вот синтезу речи письменной не уделено достаточно внимания, хотя проблемы в этой области есть очень интересные. Приходилось ли вам читать подобные фразы: «Мария поделился ссылка», «Забронировать отель в Москва», «Уважаемый(-ая) Иван Петрович», «21 файла(-ов) выбрано»? Почему компьютерам с таким трудом дается русский язык и как научить их правильно говорить по-русски? Об этом данная статья.

Все больше текстов в современном мире синтезируются компьютером. И хотя до уровня Толстого или Достоевского «искусственному разуму» еще очень далеко, генерировать тексты по составленному человеком шаблону у них получается неплохо, а главное — быстро. Количество веб-страниц, электронных писем и СМС-сообщений, генерируемых компьютерами в день, исчисляется миллиардами. В типовой текст договора компьютер мигом подставит ФИО представителя контрагента, дату, номер, сумму (цифрами и прописью), процентные ставки и сроки. Клик — и договор готов!

Все это работает здорово, пока программисту не поставят задачу собрать из частей предложение. Вот тут и начинают вылезать наружу швы и белые нитки. А все потому, что внутри предложения действуют особые грамматические законы, требующие согласования их частей по родам, числам и падежам. Проще говоря, перед подстановкой в шаблон данные бывает необходимо просклонять.

ПОДХОДЫ К РЕШЕНИЮ ПРОБЛЕМЫ

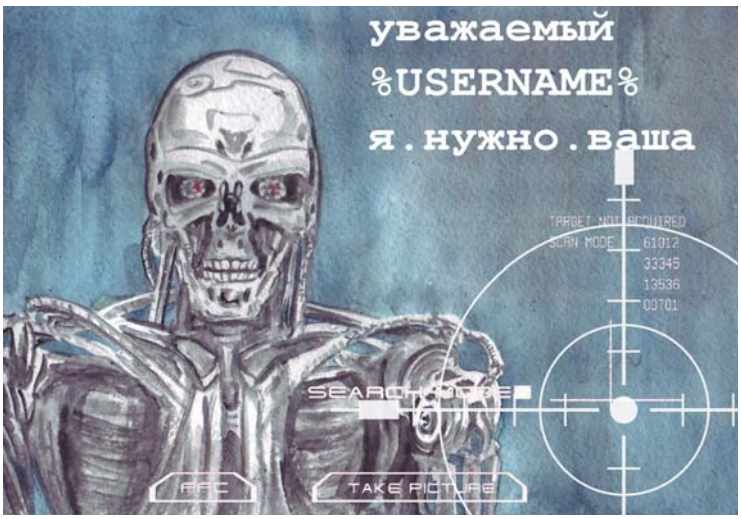
Первый вариант — просклонять вручную. Такой подход хорошо работает на малых объемах. Например, вам нужно отобразить в почтовом ящике сообщение: «У вас 5 новых писем» и варьировать эту надпись в зависимости от количества писем. Тогда вам понадобятся всего три падежно-числовые формы: «новое письмо» (для 1, 21, 31…), «новых письма» (для 2, 3, 4) и «новых писем» (для 0, 5, 6, 25…). А вот задача чуть посложнее: необходимо из названия языка (английский язык, французский язык, суахили…) построить фразы: «на английском языке», «на французском языке», «на суахили»... Официальных языков в мире не так уж много — всего 95. Можно просклонять и вручную. Еще немного усложним задачу: вы хотите отображать место рождения пользователя так: «Родился в Москве», причем место рождения пользователь вводит сам. Географических названий в мире порядка 8 млн (по версии geonames.org). «Ручной» подход здесь неприменим.

Второй — перестроить фразу так, чтобы склонение не требовалось: «Новых писем: 5». Подход вполне прагматичный, и в некоторых ситуациях подобный «уход от проблемы» является приемлемым решением. Увы, не всегда это возможно и не всегда желательно. Бывает, что текст документа закреплен нормативно или заказчик не хочет его менять. К тому же бывают и объективные причины, почему стоит использовать склонение: текст лучше читается и больше нравится поисковым системам (SEO).

Наконец, третий вариант: использовать программу склонения.

В области автоматической обработки текстов (Natural Language Processing, NLP) задачу склонения и спряжения слов решают программы, называемые «морфологическими анализаторами» (например, Mystem [1] или pymorphy [2]). Анализаторами они называются потому, что основная их задача — анализ слова, например сказать, что «кошку» — это слово «кошка» в винительном падеже единственного числа. Как правило, такие программы решают и обратную задачу — синтез слова на основе начальной формы и заданных грамматических параметров (падежа, рода, числа). Казалось бы, это то, что нам нужно? Не совсем. Дело в том, что морфанализаторы хорошо справляются с отдельными словами, а словосочетания им не по силам. А нам в общем случае нужно склонять именно словосочетания, например «Нижний Новгород» или «отделение банка». Морфанализатор просклоняет нам все подряд, включая слово «банка», и получится «в отделении банке».

Существуют ли программы для склонения целых словосочетаний?



Поиск в Google нам покажет, что таких программ существует множество. При ближайшем рассмотрении оказывается, что, как правило, они ограничены функцией склонения ФИО по падежам, так как это наиболее востребованная функция и наиболее простая в реализации. Изредка встречаются попытки автоматизировать склонение должностей и названий отделов.

Создать по-настоящему универсальный инструмент для любых слов и словосочетаний, выполняющий склонение по всем трем параметрам — род, число и падеж, — очень непростая и интересная задача.

ПРОГРАММА «МОРФЕР»

Автор данной статьи с 2003 г. работает над такой программой. Она называется «Морфер» и предназначена для склонения слов и словосочетаний. Программа внедрена на сотнях предприятий России, ближнего и дальнего зарубежья и пользуется все возрастающим спросом.

Программа реализует следующие функции:

- склонение слов и словосочетаний по падежам и числам;
- склонение прилагательных по родам;
- определение пола человека по имени;
- пропись чисел и денежных сумм;
- согласование единиц измерения с числом;
- образование прилагательных от названий городов и стран.

Функция склонения по падежам и числам — основная в программе. С ее помощью можно склонять как личные имена (ФИО), так и произвольные словосочетания, например звания, должности, названия отделов, географические названия и практически все, что склоняется в русском языке. Фамилия, имя и отчество могут подаваться на вход в любых комбинациях: Чичиков Павел Иванович; Павел Иванович Чичиков; Чичиков Павел; П. И. Чичиков и т. п.

Имя необязательно должно быть трехкомпонентным. Программа учитывает двойные имена и фамилии, а также национальные суффиксы и служебные слова:

- *Жозеф Луи Гей-Люссак*
- *Шихлинская, Нигяр Гусейн-Эфенди кызы*
- *Барклай-де-Толли, Михаил Богданович*

Имена собственные представляют для программы склонения особую проблему. В то время как фонд нарицательных слов очень хорошо представлен в имеющихся словарях, данные об именах собственных весьма отрывочны и малопригодны для машинной обработки. Большинство словарей являются орфографическими, т. е. содержат только написание имен, изредка — ударения, что тоже полезно, так как ударение иногда влияет на склонение, например:

Окончание безударно
Петро́вич – Петро́вич**ем**
Па́ша – Па́ш**ей**
Софи́я – о Софи́**и**
Окончание ударно
Ильи́ч – Ильич**о́м**
Алим-Паша́ – Алим-Паш**о́й**
Зульфия́ – о Зульфи**е́**

Однако для полного и точного склонения имен собственных в словаре должна быть соответствующая информация. После довольно продолжительных поисков был найден «Словарь собственных имен русского языка» Ф. Л. Агеенко [3] общим объемом 38 000 единиц, содержащий информацию о постановке ударения, произношении и склонении. Словарь имеет богатую историю, начинающуюся с 50-х годов прошлого века. Первоначально он существовал в виде картотеки «трудных слов» для дикторов советского радио. Два первых издания были выпущены Радиокомитетом для внутреннего пользования на правах рукописи. С 1960 по 2000 г. в государственных издательствах вышло восемь изданий Словаря, пять из них — под редакцией профессора Д. Э. Розенталя. Автор программы «Морфер» приобрел лицензию на использование словаря у издательства «Мир и Образование». Была проделана работа по приведению словаря к машинно-читаемому виду.

Определение пола человека по имени

Несмотря на то, что в современном мире определение половой принадлежности человека становится все более сложной проблемой, законы русского языка зачастую требуют от нас определить, кто перед нами: *уважаемый* или *уважаемая*, *действующий на основании* или *действующая*, *написал* комментарий или *написала*. Программа с успехом решает эту задачу благодаря встроенному словарю. Например, программа «знает», что Саят — мужское имя, а Асият — женское.

Программа «Морфер» представляет собой морфологический анализатор.
Возможные ошибки — 1,5%.
Объем программы — около 1 Мбайт.
Разработчик — Сергей Слепов.
Сайт программы — http://morpher.ru
Похожие программы: Mystem, pymorphy, Verifika и др.

Пол определяется не только по имени. Учитываются все компоненты ФИО, например:

- *Саша Иванов* (мужской);
- *Саша Иванова* (женский);
- *Карен Алиевич* (мужской);
- *Карен Алиевна* (женский).

Большинство аналогов «Морфера» требуют указания пола или отчества для правильного склонения, «Морфер», наоборот, сообщает вам пол.

Пропись чисел и денежных сумм

Часто при генерации договоров, актов, счетов и накладных требуется указать «сумму прописью», т. е. перевести *123* в *сто двадцать три*. Функция прописи работает не только с целыми, но и с дробными числами, позволяя формировать такие выражения, как *18,3% (восемнадцать целых три десятых процента) годовых*. При необходимости пропись можно поставить в нужный падеж: *в течение 21 (двадцати одного) календарного дня*.

Технические параметры программы

Программа имеет множество реализаций в виде библиотеки (подключаемого модуля) для различных языков и сред программирования, включая: C/C++ (Windows и Linux), Delphi, PHP, 1C, Excel, .NET и даже SQL Server.

Программа очень компактная, и в большинстве реализаций ее размер не превышает 1 Мбайт (для сравнения — Mystem имеет размер 15 Мбайт).

Все фунции программы потокобезопасны (thread-safe), что особенно актуально при ее использовании в серверных приложениях.

Надстройка для Excel

Реализация программы «Морфер» в виде надстройки для Excel имеет то преимущество, что вам не нужно ничего программировать, чтобы ей воспользоваться: ввод и вывод данных осуществляется посредством графического интерфейса Excel.

«Фишки»

В программе учтено очень много тонкостей, неочевидных на первый взгляд. Например:

- Склонение числительных с наращением: *1-я Парковая улица, 5-е колесо*. Учет сокращений: *о-в Святой Елены*.
- Корректная работа с буквой Ё. Там, где нужно, Ё при склонении заменяется на Е: *ёж — ежа, Пётр — Петра, копьё — о копье*. Программа старается следовать вашему стилю употребления точек: если Ё была на входе, то будет и на выходе, иначе — Е: *копье — копьем, копьё — копьём*. Но неожиданно программа может обнаружить свою «грамотность»: *щека — щёку, бедро — бёдра*. Поэтому, если вы противник Ё, заменяйте ее на Е самостоятельно.
- Исправление латинских букв. Если вы случайно в фамилии Сидоров наберете латинскую букву C, программу это не смутит, и при склонении она будет заменена на русскую С.
- Автоматический выбор предлогов о/об/обо: *о лете — об осени — обо всем хорошем*.
- Автоматический выбор предлогов в/во/на. Почему мы говорим *в Сибири*, но *на Урале? На площади*, но *в сквере? В магазине*, но *на рынке?* Ответ на этот вопрос, как и на многие вопросы в синхронной лингвистике, — «так сложилось». Можно подвести под это множество теорий, но практический результат один: запрограммировать это нельзя, можно только занести в словарь. Именно эта работа и была проделана и продолжает проделываться.
- Отдельный падеж, отвечающий на вопрос «где?». Дело в том, что в русском языке есть еще один падеж, о котором нам в школе не говорили. В специальной литературе он называется «локатив», «местный падеж» или «второй предложный». Для большинства слов он совпадает по форме с обычным предложным, но у некоторых слов отличается: *о лесе — в лесу, об аэропорте — в аэропорту, о Крыме — в Крыму*. (Вообще, вопрос о количестве падежей в русском языке очень интересный. Математически точное определение этому понятию предложили Успенский и Зализняк [4].) Поэтому для правильного склонения необходимо учитывать эту особенность таких слов. Да и предлог в этом падеже другой (*в* или *на* — см. выше).
- Склонение «по аналогии». Малоизвестные и придуманные слова программа способна просклонять по аналогии с известными ей словами. В большинстве случаев результат совпадает с ожидаемым, хотя иногда все-таки приходится объяснять программе, что *бизнес-леди* склонять не нужно, а *пицца* во множественном будет *(5) пицц*, а не *пицец*. «Объяснение» заключается в занесении слова в словарь, что является основной работой по совершенствованию программы.

Часто ли программа делает ошибки?

Тестирование 2009 г. показало 1,5% ошибок. Для тестирования были выбраны случайные фамилии, имена и отчества, из которых получилось более 1000 тестов. С тех пор тести-

рование больше не проводилось, а качество склонения только улучшилось. Почему больше не проводилось тестирование? Потому что составление тестов — весьма трудоемкий процесс. Можно ли использовать для нового тестирования уже составленные тесты? Нет, нельзя. На них программа покажет стопроцентно правильный результат — автор об этом позаботился. Проверять программу на известных тестах — все равно что сдавать экзамен, имея под рукой список правильных ответов.

Что делать, если программа неправильно «склоняет» вашего директора?

В большинстве вариантов программы имеется так называемый «пользовательский словарь», который позволяет корректировать склонение словосочетаний. Словарь представляет собой обычный XML-файл, в котором вы можете указать, как именно склонять ФИО вашего директора.

ПРИМЕРЫ ВНЕДРЕНИЯ

Программу автоматического склонения целесообразно применять, когда объем склоняемого материала велик либо когда склоняемые слова и словосочетания неизвестны на этапе разработки программы (читаются из файла, базы данных или веб-сервиса, вводятся пользователем). Экономический эффект очень сильно зависит от области применения. Вот примеры внедрения программы «Морфер»:

- Бюро переводов «ТрансЛинк СПб» работает с двуязычными текстами формата XLIFF, которые после выгрузки из программы SDL Trados Studio проверяются на предмет правильности используемой терминологии при помощи программы Verifika. Переводчики столкнулись с проблемой, что программа работает некорректно для русского языка, т. к. не учитывает падежные окончания. Проблему удалось решить при помощи «Морфера». Это экономит переводчикам массу времени, т. к. количество терминов в одном проекте может исчисляться тысячами.
- Лаборатория геоинформационных технологий ФГБУ «ААНИИ» использует «Морфер» в системе «Особо охраняемые природные территории России» для формирования кадастровых отчетов и веб-страниц системы. Склоняются географические названия, названия природных территорий и нормативных документов.
- Онлайн-сервис «Документовед» использует программу для генерации юридических и бухгалтерских документов на основе данных, вводимых пользователями.
- ООО «СКАЗКИПРО» использует «Морфер» для составления именных сказок для детей про них же самих. В сказку подставляется имя ребенка, имена его родственников, любимых игрушек и т. п. Учитывается пол ребенка.

С 2009 г. программу приобрели более 300 клиентов, среди которых: ООО «РосИнтеграция», ФГУП «ГосНИИАС», ОАО АКБ «РОСБАНК», ЗАО «РАМЭК-ВС», ЗАО ЦНТ «Парус», ООО «Медиасоюз» (Украина), АО «Цеснабанк» (Казахстан), АО «Нортал» (Эстония), Snapkeys Ltd (Израиль), Nuance Communications (США).

ОСНОВНЫЕ НАПРАВЛЕНИЯ РАЗВИТИЯ

- Пополнение словаря программы. Жизнь не стоит на месте, и в русском языке постоянно появляются новые слова — названия новых товаров, профессий, занятий: *айфон, айпад, бари́ста, отельер, мерчандайзер, кёрлинг, вейкбординг…* Имен собственных, географических названий — необъятное множество. К счастью, с недавних пор составлять словарь помогают студенты филфака НИУ ВШЭ.
- Поддержка новых платформ и языков программирования. В ближайших планах — Java.
- Охват новых естественных языков. Имеется частичная поддержка украинского: склонение ФИО и пропись чисел.

СЕРТИФИКАТЫ

- Свидетельство о регистрации программы для ЭВМ № 2012613019, выдано 28 марта 2012 г. Федеральной службой по интеллектуальной собственности («Роспатент»).
- Программа прошла сертификацию на соответствие требованиям обеспечения режима секретности 2-го уровня в составе специализированного ПО для нужд Министерства обороны Российской Федерации. ⬣

Иллюстрации Анастасии Поповой

ЛИТЕРАТУРА

1. Ilya Segalovich. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. https://tech.yandex.ru/mystem/
2. Mikhail Korobov. Morphological analyzer/inflection engine for Russian language. https://github.com/kmike/pymorphy2
3. Агеенко Ф. Л. «Словарь собственных имен русского языка. Ударение. Произношение. Словоизменение». ISBN 978-5-94666-588-9.
4. Успенский В. А. К определению падежа по А. Н. Колмогорову. Опубликовано в продолжающемся сборнике: Бюллетень Объединения по проблемам машинного перевода. М.: I МГПИИЯ, 1957. № 5.