Мир ПК

04.10.2005г.

Мир ПК №10, 2005

Переводим с ПРОМТ

Когда речь заходит об искусственном интеллекте, первым делом обычно вспоминают игры и шахматы. Поисковые машины или системы автоматизированного перевода, как правило, в расчет не берутся.

Георгий Корсаков

Когда речь заходит об искусственном интеллекте, первым делом обычно вспоминают игры и шахматы. Поисковые машины или системы автоматизированного перевода, как правило, в расчет не берутся. Между тем анализ запросов на естественном языке или перевод с одного такого языка на другой — наиболее перспективная в практическом плане область применения «думающих» машин.

Когда в начале 90-х годов появились первые версии «переводчика» Stylus, от него ждали едва ли не чуда. Велико же было разочарование, когда при переводе выражения «хотя плоть немощна, дух силен» с русского на английский и обратно в результате получилось «хотя мясо протухло, запах сильный» (Е. Козловский, 1997). Будучи свидетелями торжества мультимедийных технологий, пользователи хотели столь же эффектных программ перевода.

Сейчас основной областью применения подобных продуктов является максимально возможная автоматизация труда профессионального переводчика, а вовсе не решение проблем языкового барьера, скажем на форумах в Интернете. Насколько же десять прошедших лет улучшили положение дел с компьютерным переводом? Может ли нынешняя машина существенно уменьшить объем низкопроизводительного «ручного» труда? Ответ на этот вопрос мы попробуем найти на примере продуктов российской компании ПРОМТ — PROMT Family 7 и PROMT Translation Suite 7.

Модуль для IE в действии

Комплектация Family 7 варьируется от минималистского Internet до богатейших Professional и Premium. В состав пакета могут входить многочисленные модули, добавляющие функции перевода в популярные программы вроде Adobe Reader или Microsoft Office. Также доступна PRO-версия основного инструмента PROMT, совмещающего в себе многооконный текстовый редактор с универсальным средством настройки системы перевода. Дополнительные возможности, такие как резервирование данных или использование баз ассоциативной памяти, реализуются посредством внешних утилит. Мы рассмотрим версию Professional как наиболее сбалансированную по соотношению цена/наполнение.

Однако прежде чем приступать к последовательному знакомству с этим эффективным инструментарием, неплохо бы вкратце разобраться с принципами его работы и постановкой задач.

Все действия над проектом происходят непосредственно в редакторе PROMT. Именно

эта программа включает большинство настроек и имеет богатые возможности по импорту/экспорту/правке файлов. Редактор открывает файлы основных текстовых форматов, бинарные документы Word (*.doc), а также распознает большинство графических форматов и PDF с помощью средств встроенной системы OCR или внешнего приложения, такого как FineReader. Увы, простой импорт информации в программу может потребовать от пользователя сугубо компьютерных, а не языковых знаний.

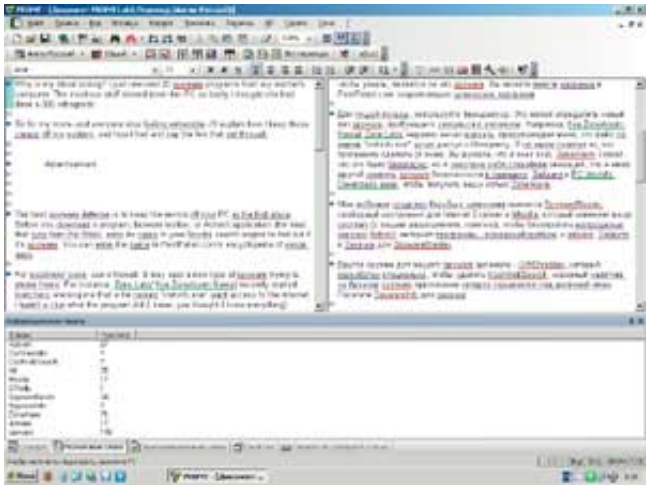
Пользователи современных текстовых процессоров, таких как Word, давным-давно не испытывают особых проблем с форматированием текста. Между тем для текстовых файлов, скажем DOS или UNIX, — а и тех и других в Интернете предостаточно — используются различные методы завершения строки. Для DOS это CR/LF, для UNIX — только LF. Таким образом, в DOS каждая строка завершается символом возврата каретки CR, который вводится нажатием и в обычном понимании обозначает конец не строки, а абзаца. Для UNIX же, наоборот, текст открывается в виде одной нечитабельной строки. Чтобы избежать проблем с расстановкой абзацев в PROMT, следует пользоваться соответствующими вариантами импорта текста: «текст с завершением строки» для DOS и «просто текст» для UNIX. В этом случае программа попытается расставить хотя бы часть абзацев автоматически. Другая «засада» может поджидать при злоупотреблении OCR: импорт 300-страничного PDF-файла занял примерно полчаса и привел к созданию более чем гигабайта временных файлов. А на его сохранение в тексте посредством Adobe Reader и последующий импорт в PROMT ушло менее 10 с. Так что не стоит искать проблем там, где их легко избежать.

Импортированный текст неизбежно будет содержать некоторое число грамматических ошибок. Каким образом это отразится на качестве перевода, вполне понятно, так что лучше сделать проверку заранее. Для этих целей можно воспользоваться либо средствами Office (если таковой присутствует), либо комплексом ORFO. Второй вариант на практике оказывается значительно удобнее. Помимо гораздо более качественной проверки нет нужды заниматься экспортом/импортом файла в Word.

Когда исходный файл подготовлен к переводу, следует определиться с порядком действий. В первую очередь он зависит от цели: необходимо ознакомиться с иностранным текстом «по диагонали» или нужен действительно качественный перевод, годный для публикации. Первый вариант мы отдельно не рассматриваем, возможности программы переводить «на лету» станут вполне очевидны в процессе доведения качества до нужного уровня. Поэтому основной упор делается на множество циклов для улучшения качества перевода и задействования максимального числа средств автоматизации.

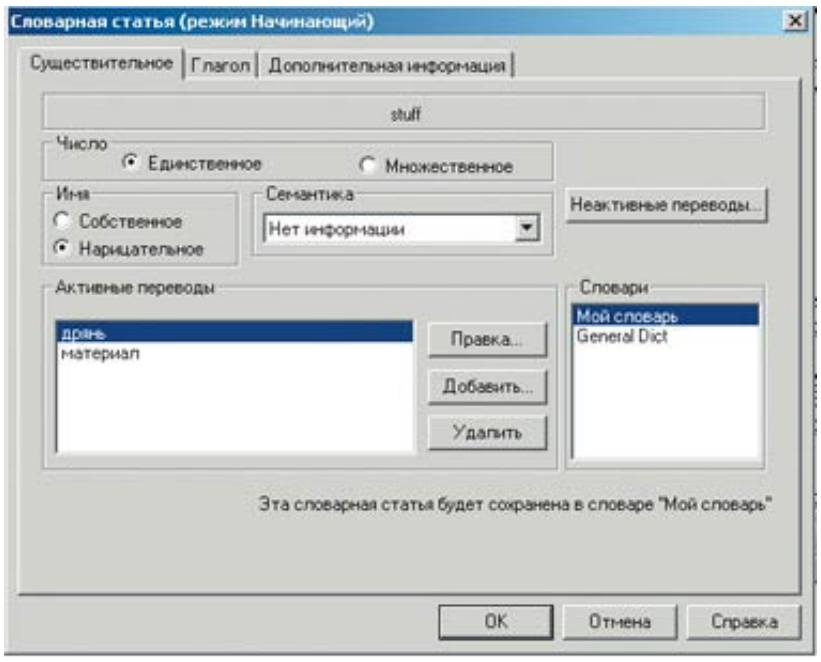
Если работа идет с небольшим текстом, вполне логично сразу перевести его целиком. На этом этапе некоторое количество слов не будет найдено в словарях. Другие же, наоборот, окажутся с несколькими вариантами перевода. Незнакомые слова могут быть как требующими перевода, так и многочисленными именами собственными, названиями и адресами. Если требуется перевод, достаточно занести слово в текущий или новый пользовательский словарь. При этом на уровне доступа «новичок» вам предложат указать лишь часть речи и перевод в исходной форме. В дальнейшем программа попробует автоматически выбрать нужное склонение. Ну а слова, не требующие перевода, можно зарезервировать и не переводя. При этом лучше указать транслитерацию, скажем, для имен собственных, чтобы Bill Gates превратился во вполне кириллического Билла Гейтса. Однако все тот же Гейтс способен присутствовать в исходном тексте как bill.gates@microsoft. com, и в этом случае транслитерация лишь

добавит нам работы руками. На это существуют препроцессоры — средства обработки текста посредством «фабричных» или пользовательских установок. Помимо этого есть возможность применить макросы Visual Basic, Jscript или Perl. Таким образом, мы имеем вполне удобные инструменты для устранения самых явных проблем.



Главное окно PROMT

Однако многовариантные слова также способны направить контекстную обработку текста по неверному пути. Работа со словарями в рамках одной тематики перевода включает в себя не только выбор нужного направления и соответствующих словарных баз, но и некую иерархию. Словари расположены по старшинству, и при выборе варианта перевода программа отдает предпочтение именно тому, который найден в более «главном» словаре. Если мы имеем дело с текстом вполне однозначной тематики, никаких сложностей эта схема не вызывает. Таким образом, перевод, скажем, лицензии на ПО в основном апеллирует к «информатике» и «юриспруденции». Самым старшим является пользовательский словарь, куда и заносятся все неоднозначные трактовки вкупе с незнакомыми словами. А вот со смешанными текстами все обстоит куда сложнее. И однозначной иерархии словарей в рамках такого текста выстроить не удастся. Можно, конечно, менять каждый некорректный вариант вручную, но не этого ли мы изначально стремились избежать? И здесь на помощь приходит абзац как основная рабочая единица. Если разбивка выполнена корректно, то смысл и тематика вполне однородны именно в рамках одного абзаца. В следующем примере рассматривается как раз такой подход. Аналогично дело обстоит и со словосочетаниями. При вводе в словарь некоего оборота программа запросит данные для каждого отдельного слова. Это позволяет надеяться, что сочетание не станет мертвым грузом в единственной форме и будет склоняться автоматически.



Добавление слова в пользовательский словарь

В качестве входного текста на английском языке взята статья автора американского журнала PC World Стива Басса «A Two-Pronged Spyware Defense», в русском переводе названная «Обороняемся от программ-шпионов». С переводом можно ознакомиться в «Мире ПК», №6/05, с оригиналом — на pcworld.com. Несхожесть названий уже ярко иллюстрирует огромную разницу между дословным переводом и художественной обработкой для публикации в журнале. Однако из многих переводных материалов журнала «Мир ПК» именно этот оказался самым подходящим для сравнения с оригиналом.

В первом же абзаце попалось словечко stuff, которое Wikipedia определяет как something means nothing («нечто не имеющее конкретного значения»). И действительно, в различном контексте значение этого слова может варьироваться от «наркотик» до «пользовательские данные». PROMT дословно перевел это как «материал». Переводчик «Мира ПК» воспользовался подходящим словом «мерзость» которое и было добавлено в словарь. В результате машина автоматически изменила «коварный материал» на «коварная мерзость» с согласованием по женскому роду. Знакомое каждому Spyware лучше зарезервировать, чем вводить громоздкий перевод. А вот «done a 386 retrograde» оказалось программе не по зубам. Художественный перевод «поменяла на древнюю 386-ю машину» имеет совершенно иную структуру, и добиться аналогичного результата от алгоритма вряд ли возможно. В данном случае оказалось проще заменить текст вручную и добавить весь фрагмент в базу АП — на случай повторной работы с текстом.

В следующем абзаце пришлось столкнуться с проблемой «американского английского». Англо-русское направление перевода подразумевает некое соответствие исходного языка литературному английскому. Для классики вроде holiday — vacation или trainers — sneakers предусмотрена пакетная конвертация. А вот с нарушением самой структуры текста и пропущенными словами уже ничего не поделать. Так, выражение «off my system» перевелось дословно как «от моей системы». Подразумевающееся автором «подальше» в тексте отсутствовало и в перевод, естественно, не попало. Фраза эта встречается в тексте несколько раз, и удобнее всего занести ее в пользовательский словарь. Если же добавить недостающее слово, то PROMT выдает корректный перевод без проблем.

Вообще, умение программы правильно подбирать склонение радует. Например,

выражение «goodness? sake», добавленное в словарь как «пущая польза», нормально вписалось в предложение «Для пущей пользы используйте брандмауэр». Абсолютно нечитабельные предложения возникают в большинстве случаев из-за неподходящего (чаще всего дословного) перевода устойчивых выражений. Например, «me included», переведенное как «я включен» вместо «включая меня». Зато достаточно было зарезервировать Zone Labs? и free ZoneAlarm, и внести в словарь слово kvetch, чтобы программа самостоятельно сгенерировала предложение: «Например, брандмауэр Zone Labs? free ZoneAlarm недавно начал ворчать, предупреждая меня, что файл по имени inetinfo.exe хочет доступ к Интернету», что почти соответствует художественному переводу в «Мире ПК».

Весь этот текст можно увидеть в *таблице*.

Красным цветом помечены два фрагмента, которые были исправлены вручную. А вот синий показывает на примере пары предложений еще одну проблему машинных переводчиков — безличные предложения. Нам с вами отлично видно, что «это» — на самом деле программа Spybot, тогда как машина никакой связи не видит и выдает предложения «про это». Впрочем, такой «интеллектуальности» пока никто и не обещал. Зато отлично видно, что достигнутый результат вполне достаточен для дальнейшей доводки в обычном текстовом редакторе без обращения к английскому оригиналу. А значит, художественной обработкой конкретно этого текста вполне может заниматься не переводчик, а, скажем, литературный редактор.

Однако на примере единственного небольшого текста адекватного представления о системе не получилось. Конечно, по сравнению с раритетными продуктами прошлого века PROMT куда эффективней. Но при этом на читабельный перевод пришлось потратить в несколько раз больше времени, чем в случае работы «по старинке», с помощью словаря и текстового редактора. И здесь наконец проявляется главное достоинство нового продукта: возможность сохранить, использовать и приумножить свои наработки.

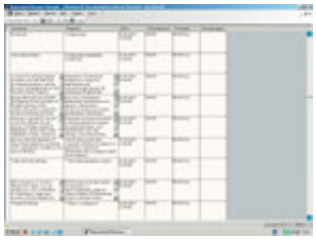
Панель инструментов PROMT в IE

Даже второй по счету материал с сайта *www.pcworld.com* был обработан несколько оперативнее. Пользовательский словарь «Моя информатика» и список зарезервированных слов послужили основой тематики перевода PCWORLD. По крайней мере такие перлы, как «особенность подавителя» (killer feature), закончили свое существование на стадии первого текста и пустого пользовательского словаря. А ведь именно такие банальности и приводят к бредовым предложениям безо всякого смысла, количество которых в рамках тематики невелико. База же АП продемонстрировала уверенное стремление превратиться с течением времени в отличного помощника.

Конечно, весь потенциал АП при наработке в несколько десятков машинописных страниц не раскроешь, но выводы сделать вполне можно. Каждое найденное совпадение избавляет от необходимости заново изобретать велосипед. Столкнувшись через полгода с тем же «done a 386 retrograde», сложно припомнить собственный вариант или удачную фразу переводчика из журнала. Да и зачем, когда есть подборка изо всех таких «велосипедов».

Таким образом, можно уверенно констатировать следующее: потенциала ядра машинного перевода вполне достаточно для корректной трансляции при условии, что все

входные данные (слова, выражения и т.д.) имеют переводные аналоги. Ну а если перевод требует недоступной машине «художественности», то база авторских интерпретаций в виде АП предоставит готовые фрагменты. В результате мы упираемся не в физическую невозможность автоматизации перевода, а в необходимость наличия тривиальной базы данных. И ее создание вполне по силам не разработчикам, а конечным пользователям.

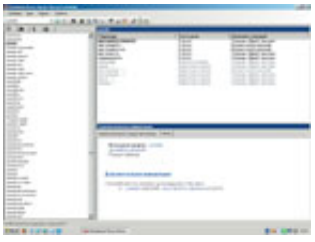


База АП

Но никакое техническое решение не выйдет за рамки хобби для энтузиастов, если им неудобно пользоваться. И в этом смысле пакету PROMT есть чем похвастаться. Взаимодействие программы и пользователя на редкость логично и предсказуемо. При возникновении проблем достаточно найти в руководстве описание проблемной функции. Да и раздела «ЧАВО» там практически нет... Зато функции программы имеют жесткую иерархическую структуру. Связь между инструментами вполне прозрачна, и грамотный пользователь быстро освоится в оболочке. Механизм работы не требуется заучивать, его нужно просто понять.

Что же касается чисто программных решений, то здесь разработчики руководствовались здравым смыслом. Многие функции в PROMT реализованы на базовом уровне или вовсе отсутствуют. Да и зачем громоздить систему проверки орфографии, если в продаже есть ORFO? «Бонус» же в виде OCR-движка бельгийской компании I.R.I.S ощутимо повысил бы стоимость продукта. Зато интеграция с наиболее популярными решениями выполнена на самом высоком уровне. В каждой поддерживаемой программе возникает панель инструментов PROMT. Все наработки в рамках переводной тематики будут доступны и в IE, и в Adobe Reader. И если с пустыми словарями и базами речи о переводе «на лету» практически не идет, то при наличии наработок по данной тематике вполне можно ознакомиться с очередным документом в онлайне. Увы, такая интеграция возможна лишь для ограниченного числа программ. А как здорово было бы иметь поддержку расширений с открытой спецификацией! Тогда пользователи смогли бы самостоятельно добавлять мощные функции перевода в любимые приложения, что лишь способствовало бы популярности продукта.

Пора, однако, познакомиться и с пакетом PROMT Translation Suite 7. В отличие от уже рассмотренного он ориентирован на более подготовленных специалистов. Вряд ли профессионалу, работающему в узкой предметной области, понадобится добавление к IE или упрощенный переводчик PromtX. А вот дополнительная функциональность будет нелишней. Именно это и предоставляет PROMT Translation Suite.



Электронный словарь и структура статьи в нем

В отличие от других продуктов семейства PROMT Family 7.0 он является «вещью в себе», никакой интеграции в рабочее окружение ОС не предусмотрено. Зато дополнением к ядру машинного перевода служит полноценная система класса Translation Memory. Концепция Suite не предполагает работы с настройками по умолчанию, напротив, каждый элемент рабочей среды, начиная от интерфейса и закачивая запросами к базе переводов, предусматривает тонкую настройку под конкретного пользователя и его специфические нужды. Пополнение базы переводов возможно не только за счет наработок в среде Suite, но и посредством импорта данных в формате TMX, являющегося независимым от конкретной системы стандартом обмена между базами переводной памяти. Таким образом, в Suite изначально предусмотрена как совместная работа с уже имеющимися системами TM, так и комфортная миграция пользователей.

В чем же потенциальные преимущества Suite перед пакетом Professional или отдельно взятыми системами TM?

Для организации полноценной работы с TM в Professional требуется дополнительно приобрести, во-первых, систему TRADOS, а во-вторых, один из двух пакетов, PROMT Expert или PREMIUM, в состав которых входит приложение PROMT for TRADOS. Если надо расширить функциональность уже имеющейся системы TRADOS машинным переводом с возможностью интеграции двух систем, то этот вариант будет весьма подходящим. Выбирая же систему «с нуля» или организуя дополнительное рабочее место, стоит приглядеться повнимательнее к Suite. Возможности подсистемы TM значительно расширены по сравнению с АП в Professional, а интеграция с ядром машинного перевода выполнена на уровне одного приложения, без «посредников», как в случае с PROMT for TRADOS. Запросы к базе переводов осуществляются с указанием минимального процента совпадения за счет использования фирменного алгоритма нечеткого поиска Fuzzy Matching. Такой подход позволяет гибко лавировать между подстановкой сегмента из базы TM и генерацией предложения силами ядра машинного перевода.

Как утверждают разработчики, подобное объединение технологий позволило усилить преимущества обеих и свести к минимуму недостатки, присущие каждой из них в отдельности.

Глоссарий

TRANSFER — вид машинного перевода, осуществляющий прямую трансляцию структуры одного языка в аналогичную структуру другого.

TM — Translation Memory — технология перевода, осуществляющая накопление базы переведенных сегментов текста и средств запроса к ней.

АП — упрощенная система класса Translation Memory в продуктах PROMT Family 7.

TRADOS — наиболее известная система TM одноименной немецкой фирмы.

Ассоциативная память

Как мы уже заметили, изобилие неоднозначных слов и устойчивых оборотов вынуждает пользователя постоянно вмешиваться в процесс перевода. Большой же пользовательский словарь лишь увеличивает количество вариантов. И здесь возникает элемент помощи свыше — система ассоциативной памяти (АП). Как можно догадаться из названия, этот механизм призван сымитировать именно то свойство человеческого мышления, которое позволяет нам не задумываясь склонять слова при переводе или, скажем, различать иронию без сопутствующего символа . Человек неосознанно проводит параллели между всем виденным им когда-то и новой, но похожей информацией. Любому алгоритму до таких способностей далеко, зато перебрать огромный объем информации на предмет поиска совпадений — это пожалуйста. Задействование АП дает возможность произвести запрос в базу уже сделанных переводов. Во многих случаях тексты сходной тематики будут содержать похожие фрагменты. Так почему же не воспользоваться уже существующим качественным переводом взамен очередного диалога машина — словарь — пользователь? Именно так и поступает PROMT. При нахождении сходного сегмента в

базе машинный перевод не производится и текст просто заменяется. И тут уже дело пользователя, соглашаться или нет. Однако вероятность годного варианта несравнимо выше — ведь совпадает не отдельное слово с кучей трактовок, а целый сегмент связного текста. Понятно, что эффективная работа возможна лишь при обширной базе АП, но ее наработка происходит уже в процессе эксплуатации машинного перевода и дополнительных усилий практически не требует.

Системы машинного перевода типа TRANSFER

Подобные системы являются на сегодняшний день наиболее развитыми. Хотя бытует мнение, что долгосрочных перспектив у данного класса переводчиков немного, более «интеллектуальные» решения пока далеки от практической (и тем паче коммерческой) реализации.

В простом случае система типа TRANSFER действует как обычный компилятор и синтезирует результирующие данные на основе анализа структуры входных. Обнаружив, скажем, фразу «I want to be a doctor», программа найдет переводные эквиваленты в словаре, определит время и выберет склонение перевода слова «doctor» как «доктором».

Однако такая реализация очень легко упирается в проблемы формального описания. Если система видит в тексте лишь один уровень, например предложение, то с усложнением структуры этого предложения возрастает и сложность его формального описания. Даже человеку легче разбить любое предложение на более простые составляющие. Если предложение сложное, то сначала выделяются образующие его простые. Затем определяются части предложения (подлежащее, сказуемое), а уже потом наступает черед определить части речи, склонения, и так вплоть до состава отдельного слова. На этом уровне структура каждого языка имеет вполне устоявшуюся строгую иерархию.

Спроецировать такие правила на несколько TRANSFER-уровней — задача вполне реальная. Машина сначала рассмотрит отдельные слова, отделит корень от окончания и получит из словаря данные о части речи и правилах склонения этих слов. Следующий уровень установит связи между словами как частями речи и членами простого предложения. Таким образом, мы получаем возможность применить полученные данные сначала к сложному предложению, а затем и к абзацу. И подобная схема способна выдать весьма читабельный текст при условии однозначности трактовки каждого найденного слова и отсутствия устоявшихся словосочетаний, которых нет в словаре. При нахождении же многозначных слов задача выбора подходящего значения ложится на пользователя. Если он не согласен с вариантом машины, весь процесс обработки предложения повторяется с учетом изменений. Аналогично и добавление устоявшегося выражения «break a leg» (дословно «сломай ногу», русский эквивалент — «ни пуха, ни пера») позволит избежать направления перевода по неверному пути с абсурдным результатом.

Макросы

Встроенных в редактор PROMT инструментов вполне достаточно для выполнения многих операций над текстом. Однако для более сложной или нестандартной обработки, как правило, требуется пользовательский сценарий. В PROMT предусмотрено подключение наиболее популярных макроязыков, таких как VB Scrip, JavaScript и Perl. Объектная

модель системы перевода предоставляет макросам широкий набор атрибутов, как сугубо текстовых (шрифт, кегль, начертание), так и специфических для самого PROMT, например незнакомое или многозначное слово.

Описанный подход позволяет сценарию не только обрабатывать статичный текст (оригинал или перевод), но и рассматривать такой текст как документ PROMT. Например, в исходном тексте могут присутствовать цитаты, выделенные курсивом, которые требуется оставить на языке оригинала. Макрос находит все слова с подобным атрибутом и резервирует их, после чего осуществляется перевод. Результаты же перевода возвращаются в виде объектов PROMT RANGES; их свойства помогут определиться с дальнейшими действиями, если таковые требуются. Например, наличие свойства VARIANTS указывает на многозначное слово, а DICT_NUMBER возвращает номер того словаря, откуда взят текущий перевод. Имея же значения этого слова, вы ищете в тексте вариант перевода, который в рамках тематики вас не устраивает, и производите замену.

Возможности применения сценариев весьма разнообразны, а поддержка широко известных макроязыков позволяет сосредоточиться на работе и не учить PROMT Script.

Постоянный URL статьи: http://www.osp.ru/pcworld/2005/10/317279/