I.A. Bolshakov, A.F. Gelbukh, S.N. Galicia Haro. *Electronic Dictionaries: For both Humans and Computers* (in Russian).
J. International Forum on Information and Documentation FID 519, ISSN 0304-9701, N 3, 1999.

Электронные словари: для людей и компьютеров

И. А. Большаков,
А. Ф. Гельбух,
С. Н. Галисия-Аро

Центр Компьютерных Исследований
Национальный Политехнический Институт
г. Мехико

{igor, gelbukh, sofia} @ pollux.cic.ipn.mx

Современные электронные словари естественных языков должны стать более универсальными для их использования людьми и компьютерами, для ссылок и обработки текстов, для текущих нужд и дальнейших лексикографических исследований. В лингвистическом плане эти словари должны стать многосвязными базами данных, сходными по содержимому с толково-комбинаторным словарем по И. Мельчуку и А.. Жолковскому или с интегральным словарем по Ю. Апресяну, но с большим упором на тезаурусные связи между статьями и с хранением общераспространенных словосочетаний. В части интерфейса словари должны обеспечивать допуск программам обработки текстов, обычному пользователю (посредством сложного интерактивного браузера) и лексикографу, чтобы добавлять новые черты и данные в постоянно растущую структуру. Проблема исследуется в приложении к словарям английского, русского и особенно испанского словарей.

Введение

В последние десятилетия параллельно существовали два различных типа естественноязыковых словарей. Словари в печатной форме были ориентированы на нужды различных читателей, т.е. на обычных образованных людей (для ссылок) и для лексикографов (для дальнейших исследований). Словари в электронной форме были, в основном, рассчитаны на автоматическую обработку текстов и не были непосредственно доступны пользователю как справочное средство. Но около десяти лет назад появились электронные словари, которые в точности повторяли печатную форму и были непосредственно ориентированы на компьютерного пользователя.

После широкого распространения компакт-дисков, не говоря уже о доступе к данным Интернета, все мыслимые ограничения на размер электронных словарей и их сложность в части демонстрации текстов на экране были сняты. Очевидной тенденцией следующего десятилетия является снижение роли печатных изданий до минимума. Широкое сообщество специалистов по вычислительным наукам может принять эту ситуацию за окончательное решение проблемы электронных словарей, когда вся информация, необходимая для автоматической обработки текстов, может заимствоваться из широкого разнообразия крупных словарей академического типа на компакт-дисках.

Но ситуация не столь оптимистична. Словари, скопированные с печатных изданий, даже академически полные, не содержат массы сведений, важных для обработки текстов, и никакая полностью автоматизированная процедура не может извлечь эту информацию из текстов, ориентированных на человека.

Между тем в большинстве своем электронные словари, составляющие базы данных для программ языковой обработки, пока тоже не совершенны, даже для решения своих собственных задач, не говоря уже о полном отсутствии в них человеко-ориентированной информации и интерфейсных средств.

В данной статье обсуждаются общие проблемы двух этих форм электронных словарей, и после выявления некоторых их недостатков предлагается возможная форма электронного словаря, по содержимому подобного [1, 2], но с большим упором на словосочетания и тезаурусные связи между словами. Предполагается, что такая форма словаря подойдет для обеих указанных задач, как содержащая достижения и академической и прикладной науки. Мы описываем три группы возможных пользователей этого универсального электронного словаря и обсуждаем некоторые требования к его частям, расчитанным на разные приложения.

Некоторые недостатки человеко-ориентированных словарей

Мотивация к более детальному исследованию разных сторон современных компьютерных словарей возникла у авторов данной статьи, когда они пытались использовать материалы доступных им электронных словарей испанского языка для определения морфологических и синтаксических свойств испанских лексем в интересах будущей системы обработки текстов.

Один из этих словарей, представленных группой Анайя [3], был доступен нам через Интернет, другой же, выпущенный Испанской Королевской Академией [4], был приобретен на компакт-диске.

Вначале предполагалось, что электронная форма этих словарей способна дать надежные материалы для тех исследователей и разработчиков, которые хотят использовать сведения о словах естественного языка для автоматической их обработки. Но чуда не произошло.

Электронная версия словаря с долгой предысторией в печатном его виде (например, [4] является 21-м изданием того же словаря в печати) просто повторяет доэлектронную версию, со всеми своими достоинствами и недостатками. Если лексикография в стране-создателе данного словаря весьма традиционна и ориентируется главным образом на лексикографов, делая максимальный упор на этимологические вопросы, то те же черты окажутся типичными и для электронных версий. Никто не позаботится о нуждах автоматической обработки, для которой все свойства языка должны быть описаны абсолютно формальным способом, даже более формально, чем если бы это было рассчитано на иностранца.

Вот несколько примеров неудовлетворенных запросов к подобным словарям относительно сведений об испанских словах графического, морфологического и синтаксического характера.

В морфологии никакой автомат не может определить, что в парах типа *lunes* Vs. *mes* или *isósceles* Vs. *inglés* первая лексема является неизменной, а вторая имеет соответственно две или четыре формы. Тем самым, в электронный словарь необходимо добавить морфологические сведения, хотя бы в виде номера соответствующего флективного класса существительных и прилагательных (В [5] предложено три таких класса для испанских существительных и шесть – для прилагательных).

Изменения в графической форме существительных и прилагательных подобных *joven* (приобретают знак ударения во множественном числе) или *inglés* (теряют знак ударения во множественном числе) могли бы быть даны в виде общего грамматического правила о сохранении ударного слога для этих частей речи. Но тогда следует дать ссылку на стандартный (абсолютно формальный и выверенный для всего языка) алгоритм деления испанских слов на слоги. Однако нам такой алгоритм не известен. В этой ситуации словарь должен быть снабжен дополнительной информацией о подобных отклонениях от простейших общих правил начертания слов.

Что же касается испанских глаголов, но ни один из словарей не указывает, например, что два омонимичных глагола *aterrar*, ‘пугать’ vs. ‘приземляться’, имеют совершенно разные морфологические парадигмы, с дифтонгизацией и без нее.

В части синтаксиса все словари испанского языка содержат помету переходности / непереходности глаголов, и это очень важная черта их морфо-синтаксического поведения. Только транзитивные глаголы могут склеиваться с местоименными клитиками (*cántale*).

Однако среди транзитивных глаголов существует не менее важная подгруппа глаголов со свойством дативности (*dar*, *dirigir*, etc.), и число приклеенных клитик для них может достигать двух (одна в дательном, а другая в винительном падеже, как в *dámelo*). Эти формы являются важной частью формирования парадигмы испанских глаголов, но без пометы дативности ни читатель ни машина не смогут их правильно синтезировать или анализировать.

Более удачная ситуация наблюдается в солидных академических словарях английского языка. Отличительной чертой английской морфологии являются нестандартные окончания множественного числа у некоторых английских существительных (*thesaurus*, *phenomenon*, *index*, ...) и нестандартная парадигма у таких глаголов как *do*, *see*, *go*, ... В словарях подобных Мэрриам-Уэбстер, имеются все необходимые сведения об этих неправильностях. Эту информацию содержат и другие хорошо разработанные словари, включая двуязычные, подобные [13]. Конечно же, все эти свойства отражены в электронных версиях тех же словарей.

Для русского языка чисто формальное представление сложной морфологии каждой флективной лексемы было дано 20 лет назад в словаре А. Зализняка [14]. Разработчики программного обеспечения и издатели электронных словарных пособий немедленно использовали этот уникальный словарь для своих нужд, и среди справочных пособий на этой основе может быть названа система «Русский Филолог» на компакт-диске [15].

Однако, как только мы попытаемся найти в академических словарях какого-либо конкретного языка нечто касающееся комбинаторных свойств слов, мы обычно не находим ничего или почти ничего. Информация о валентностях существительных, прилагательных и особенно глаголов не дается систематически ни в одном из академических словарей английского языка. Только специализированные словари типа словаря Джакендорфа содержат подобную информацию для ограниченного числа глаголов.

Аналогично, в академических словарях отсутствует информация о других важных аспектах сочетаемости слов, формализованная А. Жолковским и И. Мельчуком как лексические функции [1]. Тем самым, мы не можем узнать из них, как выразить по -

испански или по-английски такие простейшие словосочетания, как *уделять внимание* или *оказывать помощь*. Для английского некоторые сведения по данному вопросу могут быть найдены в [6], а для русского и французского существуют небольшие толково-комбинаторные словари. Но в полном объеме проблема остается не решенной ни для одного из языков.

Некоторые недостатки машинно-ориентированных словарей

Главная трудность с машинно-ориентированными словарями состоит в том же: каждый из них содержит только некоторые сведения о словах, так что для обработки информации на разных уровня языка необходимо несколько различных словарей. На первый взгляд, это не должно создавать проблем, поскольку компьютерные словари механически легко объединить друг с другом. К сожалению, все не так просто.

Во-первых, могут быть различными наборы слов в разных словарях, так что в объединенном словаре некоторые сведения о конкретном слове могут отсутствовать. За исключением небольшой ядерной части, такие словари имеют тенденцию существенно различаться по покрытию. Например, словари сходных по своим целям систем FACTOTUM® SemNet (MICRA, США) и известной в научном мире системы WordNet имеют только 20% общей лексики [16]. Когда объединяются несколько или даже много словарей, практически ни одно слово не получает той полной информации, которая представлена во всей совокупности для отдельных слов.

Более того, объединение словарей – это не такая уж прямая операция. Слияние оказывается результативным, только если правильно объединить соответствующие смыслы омонимичных и полисемичных слов. Однако как число, так и нумерация этих смыслов в разных компьютерных словарях различны, и часто нет способа распознать, какой смысл в одном словаре соответствует какому смыслу в другом.

Такое несовпадение легко продемонстрировать, сравнивая любых два больших словаря. Ниже даны статьи для одной и той же испанской вокабулы *estante* (сущ. ‘этажерка / горка / (книжная) полка / ножка (у станка и пр.)’ и прил. 'пребывающий / находящийся /оседлый'). Первая из статей заимствована из [3], а вторая – из [4], с переводом толкований на русский.

estante (в [3])

1. m. Шкаф без дверок с полками.
2. m. Полка в шкафу.
3. m. Каждая из ножек, поддерживающих каркас какой-либо машины.
4. adj. Остановившийся, неподвижный.

estante (в [4])

1. a. p. us. de *estar*. Тот, кто находится или постоянно пребывает на каком-то месте: *Pedro, ESTANTE en la corte romana*.
2. adj. Применяется к скоту, особенно к шерстному, постоянно пасущемуся в пределах юридически установленной площади, к которой он административно приписан

3. Применяется к скотоводу или хозяину стада скота.
4. Элемент меблировки с полками или досками, обычно без дверок, который служит для книг, бумаг и других вещей.
5. Полка.
6. Каждая из четырех прямых ножек, поддерживающих каркас станка, в котором снуют деревянные молотки.
7. Каждая из двух прямых ножек, на которых держится вращающаяся горизонтальная ось токарного станка.
8. *Murc.* Участник пасхальной процессии.
9. *Amér.* Каждая из деревянных устойчивых к гниению подпорок, которые, будучи вбитыми в землю, служат для поддержки каркасов городских домов в тропиках.
10. *Mar.* Палка или жердь, которая укрепляется над арматурной площадкой для прикрепления к ней такелажа корабля. Ú. m. en pl.

Даже после исключения диалектологических и терминологических вариантов, разница в числе и содержимом толкований остается весьма существенной для человека, не говоря уже о машине. Примеры таких расхождений весьма многочисленны.

В человеко-ориентированных словарях обычно даются некоторые примечания, помогающие различить смыслы, в то время как в компьютерных словарях такие примечания либо отсутствуют с самого начала, либо устраняются в процессе перевода словаря в машинную форму.

К тому же человеко-машинные словари обычно дают некоторую дополнительную информацию, подобную грамматическим справкам, таблицам единиц изменений и пр. Естественно, компьютерно-ориентированные словари таких сведений не содержат, не говоря уже о толкованиях и примерах использования слов.

В общем случае, отсутствие неформальных сведений и комментариев, когда некоторый кусок информации не может быть формализован, создает затруднения при пользовании компьютерными словарями.

К идее универсального словаря

Как мы могли видеть их предыдущих примеров, существующие машинные словари, кроме естественной неполноты информации в них, порождают проблемы, вызванные рассеянием сведений по многим источникам разных типов, с различными интерфейсами, взятыми из разных исходных словарей. Конечно, существуют (или могут быть созданы) морфологические словари, например, испанского языка в печатной форме, подобной [7]. Проблема в том, что пользователю нужно просматривать и, возможно, искать в различных источниках, чтобы выявить все сведения о конкретном слове.

Наша основная идея довольно тривиальна:

- *Компьютерный словарь должен давать всю информацию о каждом слове и о языке в целом*
- *Он должен предоставлять все возможные способы доступа и поиска этой информации.*

Под всеми возможными способами доступа мы подразумеваем, кроме прочих требований, что информация должна быть доступна как для человека-пользователя, так и для внешних программных средств. Словарь содержит настолько большой объем информации и его разработка столь дорога, что представляется неприемлемым иметь, поддерживать и использовать различные версии словарей для пользователей и для машин.

Информация в словаре должна быть представлена однородно и быть доступной в интегральной среде, такой как общий браузер для пользователя или непротиворечивый API[1] для программ, оба с мощными посковыми возможностями.

Поскольку общая информация о языка, т.е. его грамматика, может рассматриваться приложимой к любому слову языка, словарь, который дает всевозможную информацию об отдельных словах, должен давать также всю имеющуюся грамматическую информацию, с перекрестными ссылками между грамматическими таблицам и индивидуальными словами.

Более того, поскольку грамматическая информация часто имеет форму алгоритма, а не таблиц, универсальный словарь должен не только демонстрировать текст на экране и давать описание алгоритмов, но и предоставлять действующие процедуры, реализующие эти алгоритмы, различные услуги по проверке и разбору, доступные как через браузер, так и через API.

Может показаться, что речь идет просто о компакт-диске со всеми известными и еще неизвестными лингвистическими сведениями, источниками и алгоритмами, а также со всеми видами лингвистического обеспечения, от корректоров орфографии до программ перевода. Мы не хотим это утверждать определенно, хотя в дальней перспективе такое направление признаем правильным. Что же мы рекомендуем, так это, во-первых, *собрать всю уже имеющуюся информацию, включая алгоритмы, в единый интегрированный словарь* и, во-вторых, начать затем совершенствовать этот словарь путем проработки частей, в которых он очевидным образом нуждается.

Ясно, что такой словарь должен быть, по крайней мере, в первое время, механической комбинацией различных словарей, заимствованных из разных источников и поддерживаемых разными группами, поскольку едва ли можно немедленно организовать крупный проект по созданию действительно нового универсального словаря. Все такие источники должны быть объединены программой, которая анализировала бы разные форматы и собирала информацию в одну непротиророречивую базу данных со всеми необходимыми перекрестными ссылками.

Для словаря это главным образом означает сбор всей имеющейся информации по каждому слову и объединение ее в единую статью. При объединении возникает множество проблем. Наиболее очевидная проблема – это избежать повторения в статье сходных разделов.

---

[1] Application Program Interface = Прикладной Программный Интерфейс, т.е. способ доступа из других программ к данной программе или запроса на некоторые услуг с ее стороны.

Хорошо формализованной информации, такой как морфологические свойства слов, легко можно придать унифицированный вид и затем слить. Однако иные секции словарной статьи, такие как толкования, очень трудны для автоматического объединения.

Ситуация усложняется еще фактом, что операции по слиянию должны включать минимум ручного труда. Поскольку в словаре используется множество различных источников, а они обновляются в произвольные моменты группами их поддержки, словарь в целом должен будет обновляться очень часто.

По мере того, как идея универсального словаря будет становиться все более популярной и участвующие группы будут проявлять все большее согласие участвовать в совместном проекте, будет возможно разработать основные посылки, а затем и стандарты, для унификации форматов различных секций словарной статьи. Это позволит воздавать все лучшие автоматические процедуры для слияния отдельных секций.

Все эти проблемы возникают, когда объединяются словари с большим пересечением целей и содержимого. В ином случае, когда они «ортогональны» друг другу, т.е. имеют разные цели и сферы практических приложений, будет достаточно механически объединить парциальную информацию, добавляя специальные метки к каждой из частей. Например, достаточно механически объединить объяснения к английскому слову *root* 'корень' из математического и биологического словарей, с соответствующими пометами сферы употреблений при каждом из значений.

Хорошй пример идеи объединение различных источников в один программный продукт с непротиворечивым пользовательским интерфейсом дается превосходной системой словарей MultiLex (МедиаЛингва, Россия), которую мы описываем с большими подробностями в следующем разделе [8]. Однако из-за трудностей, отмеченных выше, в этой системе не было сделано попытки слить информацию для каждого слова или обеспечить способ поиска по всем словарям, имеющимся в системе.

Поскольку универсальный словарь явно будет содержать больше информации по каждому отдельному слову, чем необходимо для каждого отдельного приложения и пользователя, должны быть предусмотрены мощные и хорошо продуманные поисковые средства для любого отдельного приложения или пользователя. Объем, форма и тип информации, представленной в ответе на запрос, должны полностью подстраиваться под пользователя. Это включает возможность игнорировать или прятать некоторые (или большинство) из типов информации о слове.

Теория и практика поисковых средств ныне развиты достаточно хорошо и поставка такого поискового средства вместе со словарем абсолютно не представляет каких-либо технических проблем, так что мы не будем обсуждать эти средства в деталях. К тому же конкретные требования и спецификации на интерфейс для универсального словаря выходят за рамки данной статьи.

Что же до содержимого, универсальный словарь должен в идеале содержать следующие типы информации или некоторое их подмножество:

- Орфографическая форма (формы) слова
- Его произношение, включая варианты, если таковые существуют
- Слоговая структура слова
- Морфемная структура слова и его морфологический класс (склонения или спряжения)
- Синтаксические классификационные пометы, ключая часть речи
- Толкование слова, данное как можно более формально и с системном плане непротиворечиво. Здесь уместны специальные аллюзивные и стилистические пометы к различным толкованиям
- Отсылки к семантически связанным словам, которые в идеале влючаются в некий тезаурус или семантическую сеть. На первых порах отсылки могут ограничиться синонимическими
- Синтактико-семантическая модель управления
- Примеры использования
- Этимология
- Комбинаторные свойства, которые могут включать полный комбинаторный словарь в стиле [1, 9], и/или словарь словосочетаний, позволяющие видеть списки тех слов, вместе с которыми данное слово может быть встречено в текстах [10]
- Переводы на другие языки, в идеале, в как можно большем числе, что в простейшем случае может быть достигнуто путем механического объединения двуязычных словарей
- Другие типы имеющейся информации о данном слове.

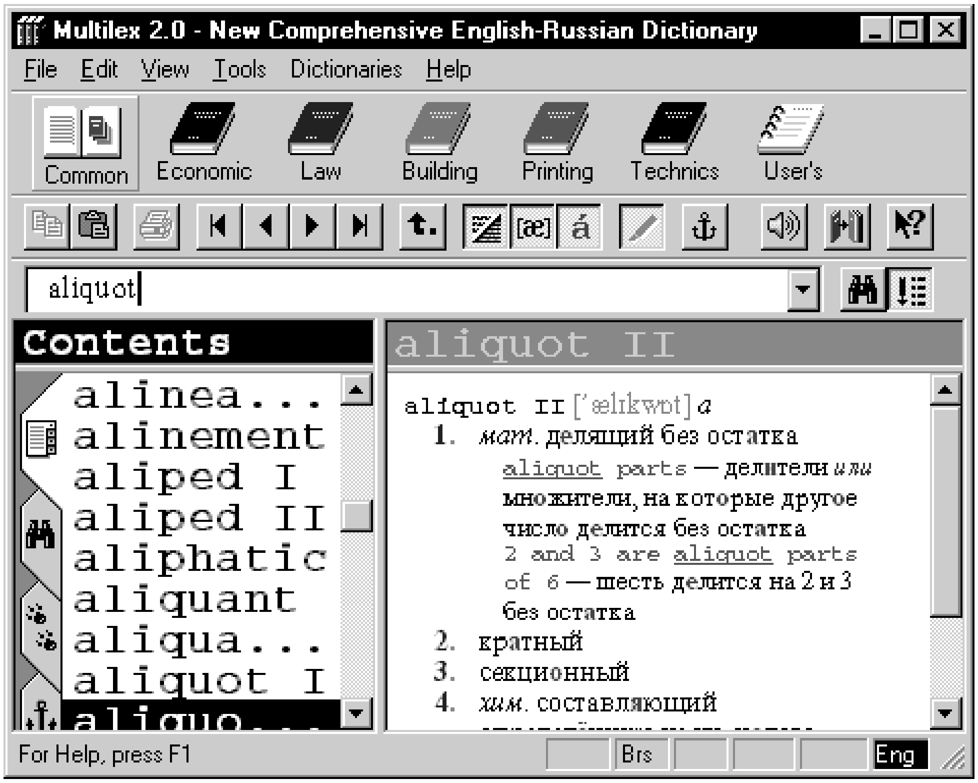
Необходимость включения в словарь комбинаторной информации должна быть подчеркнута особо, поскольку это дает очень важные сведения о словах, которые сейчас можно найти только в специальных словарях. Составить для каждого слова список слов, с которыми оно обычно и естественным образом сочетается в текстах, гораздо легче, чем дать списки лексических функций каждого слова, но в комбинации с простым тезаурусом простое перечисление оказываются очень полезным как для пользователей (при составлении текстов), так и для компьютерных программ (при синтаксическом анализе, разрешении неоднозначностей и т.п.) [10]. По нашему мнению, это и есть то направление, по которому будут развиваться современные словари в ближайшем будущем.

Ниже мы обсуждаем некоторые вопросы пользовательского интерфейса к универсальному словарю и его API.

Нужды конечного пользователя

Для обычного пользователя – человека словарь должен обеспечивать некоторый специальный интерфейс с браузером, позволяющим извлекать все необходимые знания из лингвистической базы данных и предоставлять их пользователю в понятной и удобной форме.

Примером расширяемой словарной системы, которая обеспечивает унифицированный интерфейс к разнообразию различных словарей, является отмеченная выше MultiLex 2.0. Хотя она имеет интерфейс к текстовому редактору Microsoft Word, этот продукт полностью ориентирован на пользователя. Его интерфейс показан на фиг. 1.


Multilex 2.0 - New Comprehensive English-Russian Dictionary
File
Edit
View
Tools
Dictionaries
Help
Common
Economic
Law
Building
Printing
Technics
User's
aliquot
Contents
alinea...
alinement
aliped I
aliped II
aliphatic
aliquant
aliqua...
aliquot I
aliquo...
aliquot II
For Help, press F1
Brs
Eng

aliquot II [ˈælɪkwɒt] *a*
1. *мат.* делящий без остатка
   aliquot parts — делители *или* множители, на которые другое число делится без остатка
   2 and 3 are aliquot parts of 6 — шесть делится на 2 и 3 без остатка
2. кратный
3. секционный
4. *хим.* составляющий

Фиг. 1. Русско-английский вариант системы MultiLex 2.0.

Однако авторы не объединили все имеющиеся словари в единое целое. Вместо этого пользователь может открыть, лишь один в любой данный момент, любой иной словарь из набора, показанного у верхнего края окна.

В версии 2.0 пользователь должен ввести заново то же слово, если хочет видеть его в другом словаре. Если бы другой словарь автоматически открывался на текущем слове при переключении с одного словаря на другой, это сделало бы систему MultiLex гораздо ближе и идее универсального словаря. В этом случае словарные кнопки служили бы переключателями нужной информации для текущего слова, давая пользователю иллюзию нахождения сведений о текущем слове в одном «большом» издании.

В [11] двое из соавторов программы обсуждают принципы, использованные ими при переводе существующих печатных словарей, переснятых с таких известных книг, как [13], в электронную словарную систему. Разумно повторить здесь основные тезисы этих авторов.

Она подчеркивают четыре главных различия между бумажной и электронной формой, влияющих как на способ создания словарей, так и на способ их демонстрации на компьютерном экране.

- Нет неодходимости беречь объем словаря
- Можно шире использовать средства выделения
- Можно использовать разные режимы представления информации
- Выделенные части могут быть объединены на экране

Первое различие является ключевым моментом для всех проблем создания современных электронных словарей, ориентированных на человека – пользователя, на внешнюю программу или на обоих. Для пользовательского интерфейса даже к существующим, исходно бумажным, словарям, это имеет простые, но важные следствия.

- Нет необходимости склеивать вместе в одной статье информацию о различных словах с одинаковым корнем
- Нет нужды сочленять всю информацию об одном слове в единый абзац, следует лишь визуально структурировать статью на экране
- Примеры должны использоваться значительно чаще, и один и тот же пример может быть включен в различные словарные статьи, если это необходимо.

Расчленение словарной статьи на несколько абзацев, по одному абзацу на каждую структурную единицу статьи, меняет тот способ, которым словарная статья может быть показана на экране:

- Чтобы визуально выделить различные метки и типы информации, можно использовать цвет и разные шрифтовые гарнитуры, ибо нет больше необходимости использовать их для выделения границ между структурными единицами
- Для отображения структурной иерархии статьи может использоваться вложенная иерархия абзацев (напр., части *A, I, 1., 1), a),* примеры, и т. д.
- Можно использовать разные разделители, например, тире, чтобы визуально структурировать информацию в одной строке, например, относящуюся к примерам и их переводу.

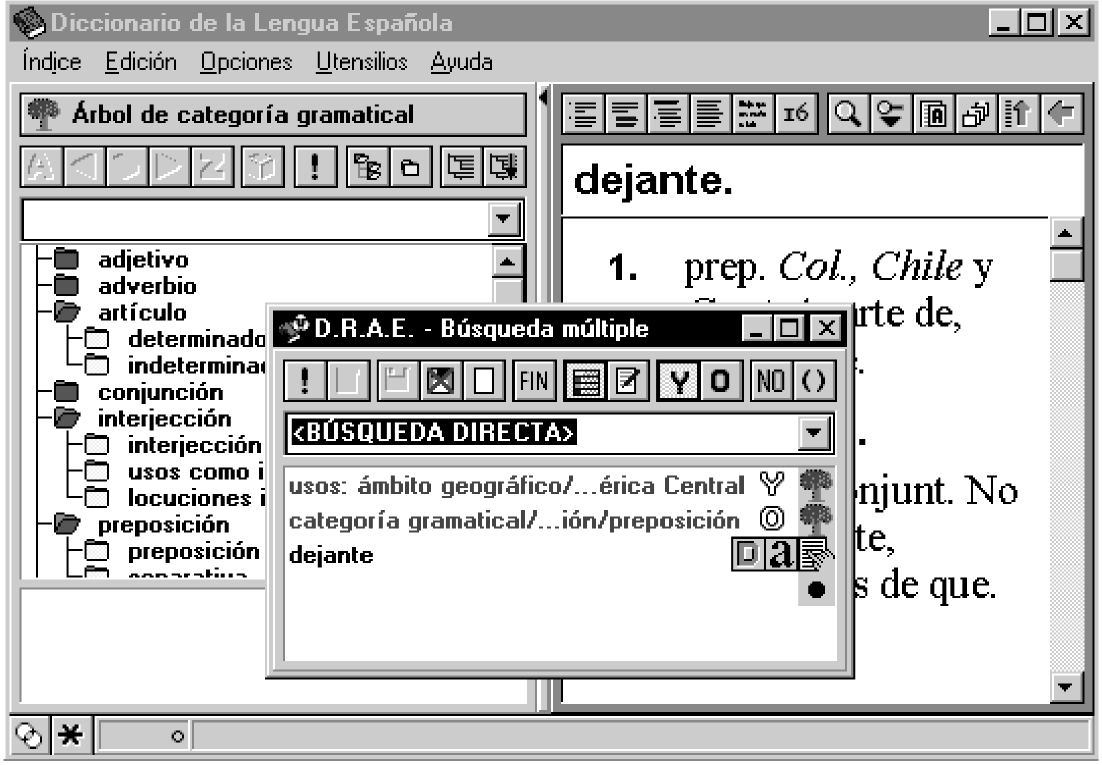
В других отношениях пользователь может использовать преимущества активной природы компьютера:

- Словарь можно адаптировать при показе, как если бы он был специализированным словарем или его независимым вариантом, например, ориентированным только на природного носителя исходного или целевого языка
- Информацию ненужного типа можно удалить с экрана или свести в иконку, например, сведения по этимологии, произношению, примеры и т.д.
- Можно представить укрупненный вид статьи или организовать словарную статью в форме дерева, позволив пользователю интерактивно расширять те ветви, в которых он действительно нуждается
- Можно подобрать информацию из различных словарных статей или различных имеющихся словарей для создания, например, списков примеров.

Однако наиболее полезно использовать в качестве реальных поискового средства активную природу компьютера, и этим можно намного превзойти его способность к автоматическому алфавитному поиску. Пример относительно мощного поискового средства дает весьма профессионально выполненный браузер для

Diccionario de la Lengua Española (Espasa Calpe, Испания). К сожалению, мы ограничены в использовании этого браузера только единственным словарем, но его поисковые средства много ближе к тем, которые могли бы подойти для универсального словаря. Отдельный экран с пользовательским запросом показан на фиг. 2.

Этот браузер позволяет пользователю видеть словарь в различных аспектах, упорядоченный по различным пометам таким как грамматические категории, исходный язык, области использования, стилистические пометы, специальные темы и многое другое. Фактически словарь может быть отсортирован по любым типам помет, используемым в нем. Эти пометы организованы в деревья, так что пользователь может увидеть только существительные или только существительные мужского рода, слова, используемые только на американском континенте или в Центральной Америке, или на Карибах, или на Кубе.


Diccionario de la Lengua Española
Índice
Edición
Opciones
Utensilios
Ayuda
Árbol de categoría gramatical
adjetivo
adverbio
artículo
determinado
indeterminad
conjunción
interjección
interjección
usos como i
locuciones i
preposición
preposición
D.R.A.E. - Búsqueda múltiple
FIN
Y
O
NO
()
16

<BÚSQUEDA DIRECTA>

usos: ámbito geográfico/...érica Central

categoría gramatical/...ión/preposición

dejante

dejante.

1. prep. *Col., Chile* y ... rte de, ... njunt. No ... te, ... s de que.

Fig. 2. Усовершенствованный поиск в испанском словаре
Diccionario de la Lengua Española.

В добавок, подобные ограничения могут быть скомбинированы в общие логические выражения, быть может, со скобками при использовании операторов AND, OR и NOT, как это показано на налагающемся окошке фиг. 2. Еще лучше то, что запрос может быть составлен путем использования удобного способа «подтащил и оставил», или, в качестве другого варианта, прямым вводом в соответствующее окошко.

Конечно, в общем случае есть много других возможных поисковых операторов для слов, включающих разные уровни их представления, таких как морфологическая нормализация, использующая морфемную или фонемную структуру, или использование регулярных выражений из букв, составляющих слово. Некоторые из таких черт, как морфологическая нормализация, воплощены в системе MultiLex, отмечавшейся выше.

В добавок, универсальный словарь мог бы обеспечить доступ к некоторым процедурам, таким как склонение или спряжение, согласование внутри словосочетаний, определение синтаксической структуры фразы, перевод фразы или текста, обучающие процедуры, игры и т.д. Отдельные из этих черты очень естественны для современных словарей и воплощены даже в некоторых карманных электронных словарях типа Spanish Master [12] фирмы Франклин, другие же, подобные способности к переводу, должны быть добавлены в универсальный словарь в будущем.

Нужды текстовых процессорных систем

Для нужд программ текстовой обработки словарь должен иметь библиотеку процедур, позволяющую обслуживать любой отдельный слой языка или всех их вместе.

Технические аспекты такого интерфейса находятся существенно за рамками данной работы. Общим соображением является то, что вся информация или максимально возможная ее часть, должна быть доступна из других программ ясным и формальным способом. Должна обеспечиваться возможность навигации по словарю, извлечения для конкретного слова его морфологических, синтаксических, семантических и комбинаторных свойств, представленных в компактной и однозначной форме.

Конкретнее, это представление не должно быть лишь строками, извлеченными из электронной версии исходного печатного словаря, со всеми противоречиями внутри одного словаря и различиями между разными словарями. Вместо этого, они должны быть членами хорошо структурированных множеств или структурами, состоящими из таких членов.

Должен существовать способ перемещения по дереву смыслов отдельного слова, от чисто тектового его представления до конкретного уникального смысла. В простейшем случае это означает, что нужную часть речи можно задавать или не задавать, группу смыслов, обозначенную через *A, I, 1., 1), a)*, можно указывать или не указывать и т.д.

В общем случае, многие из требования к визуальному представлению словарей, отмеченные в предыдущем разделе, сохраняются и для его API.

Особо важно, чтобы универсальный словарь обеспечивал услуги для других программ по разнообразным запросам, от проверки орфографии до перевода текстов, слов или фраз. Нет нужды ждать времени, когда все эти алгоритмы будут доведены до верха совершенства; словарь должен обеспечивать услуги, доступные прямо сейчас, по крайней мере связанные с морфологией.

В будущем, интерфейсы для таких услуг должны стать стандартными частями операционных систем, примерно так, как теперь делаются интерфейсы к компьютерным сетям или базам данных. Например, было бы естественно, если морфологическая нормализация стала частью стандартных поисковых срестств, а автокорректор орфографии стал частью контекстного меню стандартного редактирования, как в Microsoft Windows. Однако что для нас сейчас важно, так это что если информация не может быть обеспечена стандартными средствами, то, по крайней мере, способ доступа к ней в универсальном словаре был бы в будущем стандартизирован операционной системой.

Нужды лексикографа

Естественно, что универсальный словарь сам по себе должен быть средой, в которой лексикографы готовят новые данные для этого или иных словарей. Прекрасным примером такой интегрированной среды для лексикографа является система, построенная С. Старостиным [16].

Это означает, что словарь должен обеспечивать удобный способ модификации информации внутри себя, с внесением временных заметок и удовлетворением очень сложных запросов. Более того, словарь должен содержать свой собственный встропенный язык или интерфейс для лексикографа, чтобы он мог создавать свои собственные программы для исследования данных, присутствующих в словаре.

Информация в словаре должна снабжаться еще метками их полноты, анонимными или персонифицированными. Сведения без таких меток, т.е. потенциально неполные или не вполне ненадежные, будут доступны только привилегированным пользователям (лексикографам и системным менеджерам или иным пользователям, но в особом режиме) для дальнейших пополнений и коррекции.

Одной из частей универсального словаря, очень важной для обеих категорий пользователей, т.е., для обычных пользователей и лексикографов, являются примеры использования слов. Примеры для лексикографов должны даваться в более широком контексте, быть размеченными (в части морфологии, синтаксиса и пр.) и доступными через встроенный язык для детального изучения. Текстовой корпус также мог бы быть очень полезной частью универсального словаря, будучи нужным образом размеченным с однозначными ссылками к словарным статьям.

Заключение

Главная проблема существующих словарей, если не касаться естественной неполноты информации в них, является рассеяние различных сведений и функций по многим различным словарям и программам. Имеются словари математические и биологические, словари для человека-пользователя и для компьютера, словари морфологические, этимологические и пр. Существуют программы для автоматической проверки текста, автоматического перевода, обучения языку и индексирования, и каждая пользуется своим собственным словарем. В печатной форме это разнообразие было и остается неизбежным, но в электронной форме это нонсенс.

Мы предлагаем объединить все наличествующие типы лингвистической информации о каждом отдельном слове в единый универсальный электронный словарь. На первое время это будет механическая комбинация существующих словарей. Объединенный словарь должен обеспечить широкое разнообразие способов доступа к имеющейся информации, как для других программных средств, так и для человека-пользователя, и мощные поисковые возможности. Он должен использоваться как универсальный источник сведений для решения задач обработки текста, обучения языку, перевода и т.д.

В будущем такой словарь будет пополнен сведениями дополнительных типов, особенно, комбинаторной информацией в форме лексических функций и других словосочетаний.

Литература

1. Mel'čuk, I. A. and A. K. Zholkovsky. *Explanatory combinatorial dictionary*, in: M.W. Even (ed.), Relational Models of the Lexicon: Representing knowledge in Semantic Networks. Cambridge: Cambridge University Press, 1988, pp. 41-74.
2. Апресян Ю. Д. *Об интегральном словаре*. Семиотика и информатика, N. 32, М.: ВИНИТИ, 1988.
3. *Diccionario del Español contemporaneo*, grupo ANAYA, http: // www. anaya. es.
4. *Diccionario de la lengua Española*. Real Académia Española, Edición en CD-ROM, 1996.
5. Bolshakov, I. A. *Modelo morpfólogico formal de sustantivos y adjetivos en español*. Computación y Systemas, N 1, 1997
6. Benson, Benson, and Ilson. *The BBI Combinatory Dictionary*. M.: Русский язык, 1989
7. García-Pelayo, R., G. M. Durand, *Práctico Larousse de la conjugación*. Ediciones Larousse, México: 1983.
8. *MultiLex*, MediaLingva, Inc., http: //www. medialingva. ru /english / russian / MultilexOnline / MOLFrames.htm
9. Steele, J., ed. *Meaning – Text Theory. Linguistics, lexicography, and implications.* University of Ottawa Press, 1990.
10. Bolshakov, I. A. Multifunction thesaurus for Russian word processing. Proceedings of 4th Conference on Applied Natural language Processing, Stuttgart, 13-15 October, 1994.
11. Волович, М. М., К. П. Зоркий. *Словарь в книге и на экране компьютера* // Труды Международного семинара Диалог'97 по компьютерной лингвистике и ее приложениям, Москва, 1997.
12. Franklin Inc., BookMan, http: // www. franklin. com / products / newprod. html
13. Апресян Ю. Д. *Новый англо-русский словарь.* М.: Русский язык, 1994.
14. Зализняк А. А. *Грамматический словарь русского языка*. М.: Русский язык, 1974.
15. Русский Филолог, Агама, Россия. http: // russia.agama.com / Rfil. htm
16. Перцов, Н. В., С. А. Старостин. *О лексикографической справочной информацонной системе ЛЕКСИС по русскому языку* // Труды Международного семинара по компьютерной лингвистике и ее приложениям Диалог'95, Казань, 1995.

17. Большаков, И.А., П. Дж. Кассиди, А.Ф. Гельбух. *Русский Роже: параллельные русский и английский варианты иерархического тезауруса семантическими связями на базе пополненного тезауруса Роже* // Труды Международного семинара по компьютерной лингвистике и ее приложениям Диалог'95, Казань, 1995.