

# Электронные словари и компьютерная лексикография

Владимир Селегей

## Введение

Термин "электронный словарь" стал уже привычным. При этом атрибут "электронный" характеризует свой объект настолько же поверхностно, насколько противоположный ему атрибут "бумажный" - традиционные словари. Обычно подразумевается, что словарь на компьютере - это введенный в него бумажный словарь, снабженный удобными средствами поиска и отображения. То есть, создатели электронных словарей переливают старое доброе лексикографическое вино в новые электронные мехи. Компьютерная лексикография как область прикладной лингвистики, производящая такие словари, оказывается лишенной собственного языкового предмета. На ее долю оставляется только эффектная демонстрация канонического содержания.

Мы бы хотели предложить другую точку зрения, согласно которой компьютерная лексикография является особым направлением в практической лексикографии со своими собственными подходами не только к отображению, но и к содержанию словаря. Мы полагаем, что электронный словарь - это особый лексикографический объект, в котором могут быть реализованы и введены в обращение многие продуктивные идеи, не востребованные по разным причинам в бумажных словарях.

Необходимо сразу оговориться, что речь идет о тенденциях, потенциальных возможностях компьютерной лексикографии, часть которых еще не реализована или даже еще не осознана.

Прежде чем перейти к обсуждению новых возможностей, остановимся на проблемах классической "бумажной" лексикографии.

## Антиномии бумажной лексикографии

Плоды традиционной практической лексикографии страдают от трех фундаментальных противоречий, характерных для этой области человеческой деятельности:

1. Чем больше объем словаря, чем полнее и доказательнее описание лексических значений, тем сложнее им пользоваться.

Это противоречие привело к поляризации рынка бумажных словарей: имеется большая группа массовых изданий, довольно примитивных, но относительно удобных, которой противостоят единичные пудовые профессиональные издания, непригодные для быстрого получения информации. Характерный пример - 20-томный Оксфордский словарь.

2. Чем полнее и глубже описание лексических значений, тем в меньшей степени словарь соответствует текущей языковой и культурной ситуации.

Чрезвычайно долгий цикл создания и модификации фундаментальных бумажных словарей приводит к тому, что образ мира, который они фиксируют в системе своих значений, примеров и переводов, уже заметно отличается от действительности. Многие словари, основной корпус статей которых сформировался в языковой атмосфере середины века, представляют собой лексикографические музеи (а то и терминологические кладбища, если говорить о специализированных словарях).

3. Чем интереснее собственно лексикографическая концепция словаря, чем интегральнее средства описания лексических значений, тем уже его лексическая база.

В результате, универсальные бумажные словари демонстрируют печальное отсутствие влияния достижений теоретической лексикографии на лексикографическую практику. Научные лексикографические проекты существуют, но реализуются в виде словарей, не

покрывающих и 10% всего лексикографического пространства. Например, в знаменитом Толково-Комбинаторном Словаре (ТКС) под ред. Мельчука и Жолковского [1] описано всего около 400 лексических значений русского языка.

### **Возможности компьютерной лексикографии**

Компьютерная реализация бумажного словаря сама по себе позволяет преодолеть часть указанных проблем. К новым возможностям электронного словаря относятся:

1. Существенно более изощренные возможности показа содержания словарной статьи, включая возможность частичного показа по разным критериям ( различные "проекции" словаря), разнообразные графические средства, которые не используются в обычных словарях.

2. Использование для доступа к содержанию различных лингвистических технологий, таких как морфологический и синтаксический анализ, полнотекстовый поиск, распознавание и синтез звука и т.п.

С точки зрения пользователя смысл реализации в электронном словаре всех этих технологий состоит в том, что становится возможным быстро получить информацию, которая содержится где-то в недрах словаря и непосредственно отвечает тому запросу, который сформулирован пользователем в удобной для него форме. При традиционном подходе минимальной единицей доступа является лексема (имя словарной статьи): мы должны прочитать всю статью, чтобы определить, содержится ли в ней ответ на наш запрос. Для таких словарей, как оксфордский, это представляет серьезную проблему. Например, глагол *set* имеет там 400 только основных значений (и у многих из них имеются подзначения).

Пользователь хотел бы, чтобы словарь максимально локализовал релевантную информацию. При этом речь не идет об автоматическом выборе переводного эквивалента (если мы говорим о переводном словаре). Специфика словарного ответа в том, что он дает весьма разнообразную информацию о слове или словосочетании, а не просто переводное соответствие, предполагает активный выбор пользователя из нескольких возможных хорошо обоснованных альтернатив. Однако, попытка решить проблему адекватной реакции словаря на запрос неизбежно наталкивается на сопротивление самого словарного материала, перенесенного из бумажного словаря.

### **Новое противоречие**

Итак, мы видим новое противоречие: между новыми языковыми компьютерными технологиями и старым традиционным словарным содержанием, не позволяющим воспользоваться этими технологиями в полном объеме. Иными словами, новые механизмы требуют нового вина!

Источник этого противоречия тоже ясен: словарь представляет собой модель языка, устроенную на совершенно иных принципах, чем те формальные модели, которые лежат в основе этих технологий. И если в области морфологии противоречие еще не очень существенно, то в области синтаксиса и семантики оно становится почти непреодолимым.

Действительно, технология морфологического анализа всего лишь позволяет установить соответствие между исходной формой слова из текста и множеством лексем (словарных входов), для которой такая форма возможна. Синтаксический анализ позволяет сделать то же самое для словосочетаний, являющихся отдельными словарными входами. Однако, для всех этих технологий само словарное содержание является "непрозрачным", полностью ими игнорируется. Заглянуть "внутрь" словарной статьи позволяет только полнотекстовый поиск. Однако, этот мощный инструмент работает со словарным содержанием как с текстом на естественном языке, что резко ограничивает его возможности. Первый и очевидный шаг, на который уже идут создатели электронных словарей, это

первичная разметка словарной статьи, формализация той внутренней структуры, которая в той или иной мере имеется в хороших бумажных словарях. В результате полнотекстовый поиск может различать, к примеру, переводы, примеры использования и комментарии, что принципиально усиливает его возможности с точки зрения пользователя.

Однако, все эти меры являются поверхностными. Ясно, что задача состоит в том, чтобы единицей описания было отдельное лексическое значение, и технологии анализа могли бы устанавливать соответствие между исходным запросом и теми лексическими значениями, которые релевантны для этого запроса по синтаксическим и семантическим критериям.

В качестве иллюстрирующего примера можно привести практически любой глагол, принадлежащий ядру языка. Например, глагол "развести" может встретиться в таких контекстах:

*разводить руками;*

*разводить спирт водой;*

*разводить супругов;*

*разводить мосты;*

*разводить баранов;*

*разводить дерущихся;*

*разводить пилу;*

*разводить/разбивать сады*

*(английские эквиваленты: bring; conduct; part, separate; mix; dissolve; divorce; breed; plant, etc...)*

Задача создания такого словарного содержания, которое позволило бы сделать единицей анализа отдельное лексическое значение, а не морфологическую лексему, видится нам наиболее перспективным направлением в компьютерной лексикографии. Ясно, что для ее решения требуется "синхронизация" словарных описаний и формальных моделей, используемых технологиями анализа. В пределе это должно быть единое интегральное лексико-синтактико-семантическое описание.

#### **Читатели и писатели**

Интегральный подход к лексическим описаниям позволяет также решить и проблему "монофункциональности" бумажных словарей. К примеру, особенностью большинства бумажных переводных словарей является ориентация описания структуры лексического значения в исходном языке на лексическую систему языка перевода и на реализацию ровно одной функции - собственно перевода с языка А на язык Б в предположении, что язык А является иностранным, а язык Б - родным. Нечего и говорить, что такое ограничение делает словарь исключительно неудобным при необходимости перехода от пользовательской модели Читатель к модели Писатель.

Фактически сегодня такие модели реализуются разными типами словарей, что достаточно неудобно читателю. Поэтому интегральный подход к лексическим описаниям оправдан не только методически (и, что немаловажно, экономически), но и с точки зрения учета интересов пользователя.

#### **Проблема актуальности**

Коснемся проблемы актуальности словарного содержания.

Как уже указывалось, фундаментальные (лучшие!) бумажные словари - неизбежно словари устаревшие.

Особенно это характерно для разговорной лексики, в частности, ненормативной. В этой области отечественные классические словари предстают не только устаревшими, но попросту ханжескими.

Функции фиксации текущего состояния языка берут на себя растущие, как грибы после

дождя, небольшие словарики, обычно весьма конъюнктурные и поверхностные. Новые значения в них оторваны от своих языковых корней, плохо или произвольно объяснены.

Для массовых программных продуктов, каковыми являются электронные словари, характерны частая смена версий и наличие постоянной обратной связи с тысячами пользователей. Поэтому компьютерная лексикография - это неизбежно актуальная лексикография.

Жизнь электронного словаря должна быть похожа на нелегкую жизнь других программных систем: с маниакальным стремлением особо вредных пользователей обнаружить очередную ошибку или лагуну, и, с другой стороны, с возможностью и необходимостью поправить дело сейчас, а не через десятилетия.

Такой подход всего лишь фиксирует естественное положение дел: коллективное авторство на словарное содержание принадлежит всем носителям языка, задача лексикографа - фиксация языковых фактов и их методически правильное описание.

### **Соответствие уровню достижений лингвистической науки**

Отрыв лексикографической теории от лексикографической практики велик. Это должно быть особенно обидно для российской лингвистической науки, в которой лексическая семантика занимает особое место. Достаточно назвать такие имена, как Мельчук, Апресян, Падучева и многие другие.

Разумеется, существуют особые "концептуальные" словари, в которых лексика представлена интегрально и систематически. Например, уже упоминавшийся ТКС, созданный в рамках теории Смысл-Текст Мельчука, или толковые и синонимические словари группы Апресяна.

При этом в массовых бумажных словарях никаких следов этих идей вы не обнаружите. И именно в развитии этих идей мы видим будущее практической компьютерной лексикографии.

В этой статье мы не можем подробно анализировать теоретические концепции, являющиеся одновременно и практически полезными. Укажем лишь на следующие:

- Понятие "лексической функции", позволяющее систематически описывать несвободную сочетаемость слов. Например, то, что "войну ведут", а "экзамен - держат", что "теории выдвигают", а "мысли подают" и т.п.
- Описание семантики и практической реализации грамматического словоизменения и словообразования. Каждый язык имеет свои собственные способы грамматического кодирования смысла. И эти способы никогда не описываются в массовых словарях систематически. Например, как передать по-английски смысл "довыпендриваться", даже если знаешь как передать "выпендриваться"?
- Синтаксические описания. Здесь ситуация наиболее печальна, поскольку в массовых словарях не существует даже системы понятий, с помощью которой синтаксическая информация могла бы быть доведена до обычного читателя. Идея, что за составление предложения ответственна грамматика, изложенная в справочнике, а словарь обеспечивает перевод отдельных слов, не выдерживает критики с точки зрения современных представлений о центральной роли слова в синтаксисе.

Выход из этой печальной ситуации уже указан. Будущее лексикографии за интегральными словарными описаниями, основанными на формальных моделях, учитывающих упомянутые научные результаты. На этих же моделях будут основываться технологии доступа к словарному содержанию.

### **Заключение**

Пару лет назад накануне 1 апреля подписчикам сетевой лексикографической конференции было разослано следующее сообщение, которое мы оставим без комментариев:

*Поскольку число слов в английском языке продолжает увеличиваться, словари становятся все толще, а издательские издержки стремительно растут - лексикографы всего мира предложили революционное решение этой проблемы. На очередной встрече Ассоциации Творческих Лексикографов, ее члены единогласно проголосовали за 15-процентное сокращение всех словарей. Сокращение будет произведено пропорционально для всех букв и на всех уровнях словаря. Таким образом, к 2002 году каждый вновь выходящий словарь, от школьных до академических, будет урезан на 15% Президент Ассоциации Харли Лайкли определил это решение как "экологически корректное", указав, что меньшие по объему словари сберегают леса. Дело за малым: определить, что именно следует выкинуть*