Преобразование структуры словарной статьи Викисловаря в таблицы и отношения реляционной базы данных[1]

Крижановский Андрей

Санкт-Петербургский институт информатики и автоматизации РАН
Санкт-Петербург, 14 линия д.39, 199178
+7 (812) 328-80-71

andrew dot krizhanovsky@gmail.com

АННОТАЦИЯ

В статье обсуждается вопрос автоматического извлечения данных из Викисловаря – многоязычного многофункционального словаря, создающегося силами энтузиастов со всего мира на тех же принципах, на которых успешно работает энциклопедия Википедия. С точки зрения компьютерной обработки текста словарная статья Викисловаря представляет собой обычный текст. Руководство Викисловаря описывает структуру словарной статьи и ряд правил, которых должны придерживаться редакторы словаря. Эта структура и правила позволяют взглянуть на словарную статью с точки зрения объектно-ориентированного программирования. В этом случае сама статья и её разделы и подразделы будут соответствовать классам, а наличие каких-либо подразделов в разделах указывает на наличие отношений между классами-подразделами и классами-разделами. Такое соответствие позволяет перевести "плоский" текст Викисловаря в объектно-ориентированную форму, а именно: на основе данных Викисловаря создать экземпляры классов, присвоить значения свойствам объектов. Естественным результатом будет создание программного интерфейса (API) для работы с объектами этих классов, а по сути – с данными Викисловаря. С другой стороны, для удобной компьютерной обработки данные Викисловаря должны храниться в базе данных. В данной работе представлено, как при создании машинно-читаемого Викисловаря были решена задача преобразования структуры словарной статьи Викисловаря в таблицы и отношения реляционной базы данных, т.е. «плоский» текст словарных статей Викисловаря был преобразован и сохранён в специально разработанную реляционную базу данных. Созданный машинно-читаемый словарь содержит толкования слов, семантические отношения и переводы, извлечённые из Английского и Русского Викисловарей. Разработанное программное обеспечение находится в свободном доступе с открытой лицензией (http://code.google.com/p/wikokit) с тем, чтобы привлечь учёных и программистов к использованию построенного машинного словаря и развитию парсера.

Ключевые слова: Викисловарь, словарь, тезаурус, лексикография, машинно-читаемый словарь, парсер.

1. ВВЕДЕНИЕ

Викисловарь – это уникальный ресурс, который востребован во многих NLP задачах. Однако напрямую он использоваться не может. Он должен быть преобразован в формат, удобный для машинной обработки, т.е. в машинно-читаемый словарь (MRD).

Дело в том, что с точки зрения компьютерной программы, пытающейся напрямую работать со статьёй, словарная статья Викисловаря представляет собой обычный текст. Структура статьи, описанная в правилах Викисловаря, существует только в голове редактора, и нет никаких специальных программных механизмов, чтобы контролировать ввод данных. Отсутствие таких ограничений удобно для экспериментов и для разработки участниками словаря новых, более совершенных и удобных правил редактирования. Однако такая гибкость формата исходного текстового материала оборачивается большей сложностью при разработке программы по извлечению данных из Викисловаря, т.е. парсера. Создание такого парсера, вероятно, в чём-то сходно разработке браузера, который должен корректно вести себя при самой невообразимой HTML-разметке и при этом пытаться отобразить пользователю максимум той информации, которая поддаётся распознаванию.

Особенность Викисловаря в том, что его создаёт сообщество энтузиастов, причём далеко не все из них являются профессиональными лингвистами лексикографами. Структура словаря постепенно, но постоянно меняется, т.к. сообщество постоянно обсуждает и вырабатывает новые правила оформления статей, может изменяться структура статей, не говоря уже о том, что постоянно растёт сам словарь, в него добавляются новые словарные статьи, и более того – добавляются новые языки. Сейчас в Английском Викисловаре даны переводы и представлены словарные

[1] See English version of this paper: http://arxiv.org/abs/1011.1368

статьи примерно на 760 языках, из которых парсер распознаёт языковые коды 169 языков, в Русском Викисловаре – 343 языка, где парсеру понятны коды 337 языков.

В предыдущей работе [5] было рассмотрено решение задачи извлечения данных из Викисловаря с точки зрения разработки компьютерной программы, а именно: обсуждались требования к парсеру и архитектура программы. А в этой статье больше внимания будет уделено именно данным, т.е. будет последовательно представлена взаимосвязь структуры словарной статьи и структуры базы данных машинно-читаемого словаря.

В следующей главе описана предыстория данного проекта и текущие исследования в области построения машинно-читаемых словарей, тезаурусов, обработки данных, в том числе Викисловаря и Википедии.

В третьей главе представлено соответствие структуры словарной статьи и таблиц базы данных машинно-читаемого словаря, которую заполняет парсер. Завершает статью обсуждение результатов и оценка перспективных направлений работ, связанных с развитием данного проекта.

2. ПРЕДЫСТОРИЯ ПРОЕКТА И ТЕКУЩЕЕ СОСТОЯНИЕ ДЕЛ В ОБЛАСТИ ПОСТРОЕНИЯ МАШИННО-ЧИТАЕМЫХ СЛОВАРЕЙ

До создания парсера Викисловаря уже были получены навыки при разработке компьютерных программ для поиска семантически близких слов в статьях Википедии [8], для построения индексной базы данных по текстам Википедии [7]. Закономерное желание использовать наработанный программный код (например, набор функций на языке Java для извлечения текстов из базы данных MediaWiki) привело к следующим требования при разработке парсера:

- программный код пишется на языке Java;
- для обработки данных используется дамп базы данных Викисловаря, загружаемый в СУБД MySQL.

Несомненно, задача преобразования бумажных и электронных словарей в машинно-читаемый формат стояла задолго до появления Викисловарей [9], [15]. Однако только сейчас появился такой удивительный ресурс, предоставляющий беспрецедентный объём лексикографических данных на всех языках мира.

При решении задач автоматической обработки текста огромной популярностью пользуются тезаурусы, созданные вручную (например, WordNet). В практических приложениях также активно пользуются тезаурусы, сгенерированные автоматически, например по данным Википедии или Веб.

Не только Викисловарь, но и Википедию можно рассматривать, как тезаурус. Разрабатываются специальные алгоритмы для извлечения семантических отношений из Википедии. Например, из Японской Википедии [13] извлекают гипонимы и гиперонимы. Из текстов статей биологической тематики Английской Википедии извлекают и другие таксономические отношения для построения онтологий [4].

Тем не менее, исследований, посвящённых непосредственно Викисловарю, крайне мало. В качестве примера можно привести работу [16], в которой представлен программный интерфейс (API) к Википедии и Викисловарю (немецкая и английская редакция викисловарей).

Есть ряд работ, посвящённых сравнению Викисловарей и других тезаурусов. В нашей предыдущей работе [6] мы сравнили поиск семантически близких слов на основе Русского Викисловаря и WordNet в пользу WordNet. В работе [10] выполнено сравнение трёх ресурсов: Викисловарь, OpenThesaurus, and GermaNet. Оказалось, что Немецкий Викисловарь содержит меньше всего семантических отношений (157 тысяч на июнь 2009).

Викисловарь, в свою очередь, может быть источником данных для построения других тезаурусов. Так в работе [3] описано построение французского и словенского WordNet'ов на основе данных, извлечённых из французского, словенского и английского викисловарей.

3. СООТВЕТСТВИЕ СТРУКТУРЫ СЛОВАРНОЙ СТАТЬИ И МАШИННО-ЧИТАЕМОГО СЛОВАРЯ

Структура словарной статьи Викисловаря достаточно жёстко и однозначно задаётся правилами. Такие правила есть и в Английском Викисловаре, [2] и в Русском Викисловаре, [3] и, по-видимому, в остальных 168 викисловарях.[4]

Следует отметить, что Викисловарь (а так же и Википедия) обслуживается программной оболочкой MediaWiki, которая никак не учитывает данную

---

[2] См. http://en.wiktionary.org/wiki/Wiktionary:ELE

[3] См. http://ru.wiktionary.org/wiki/Викисловарь:Правила_оформления_статей

[4] См. http://meta.wikimedia.org/wiki/Wiktionary/Table

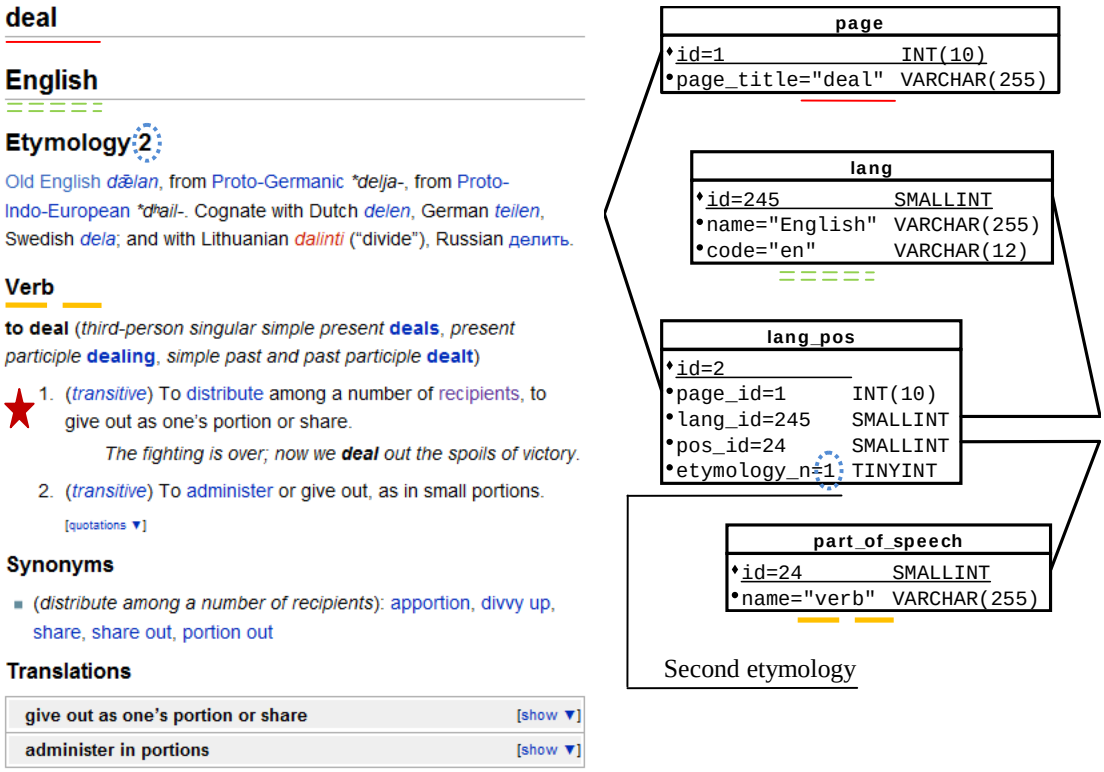

Рис. 1. Отображение данных части словарной статьи «deal» из Английского Викисловаря (слева) в значения полей таблиц машинно-читаемого словаря (справа), а именно: название словарной статьи (*deal*), язык (*английский*), частеречная принадлежность (*глагол*)

структуру словарных статей. Т.е. в базе данных MediaWiki, грубо говоря, словарной статье соответствуют только два поля: это название статьи и текст статьи в вики-разметке.

Таким образом, наличие структуры и правил форматирования словарных статей позволяет взглянуть на словарную статью как на интереснейший объект с точки зрения автоматического извлечения данных, например с помощью регулярных выражений [1]. Такое автоматическое извлечение позволит преобразовать «неявную» структуру, т.е. структуру понятную только читателю словаря, в явную, «понятную» компьютерным программам форму, чтобы обеспечить в дальнейшем успешное использование данных Викисловаря в различные проектах, связанных с обработкой текста.

Далее будет представлены фрагменты словарной статьи. Будет показано, каких таблиц машинно-читаемого словаря достаточно, чтобы сохранить данные, извлекаемые из Викисловаря.

Для удобства и наглядности на рис. 1 (слева) представлена только часть словарной статьи *deal* из Английского Викисловаря.[5]

С помощью сплошных и пунктирных линий на рисунке показано, что каждому элементу словарной статьи (слева) соответствует поле одной из таблиц (справа). Таким образом, структура машинно-читаемого словаря максимально соответствует структуре исходных данных, т.е. структуре словарной статьи Викисловаря. Итак, справа на рис. 1 представлено несколько взаимосвязанных таблиц машинно-читаемого словаря, а именно:

- *page* – это ключевая таблица в базе, содержит уникальный идентификатор (поле *id*) и имя словарной статьи (поле *page_title*), подчёркнутое сплошной линией;
- *lang_pos* – вторая важнейшая таблица. Является указателем на часть статьи, однозначно определяемой тремя параметрами:

[5] См. полную словарную статью по адресу: http://en.wiktionary.org/wiki/deal

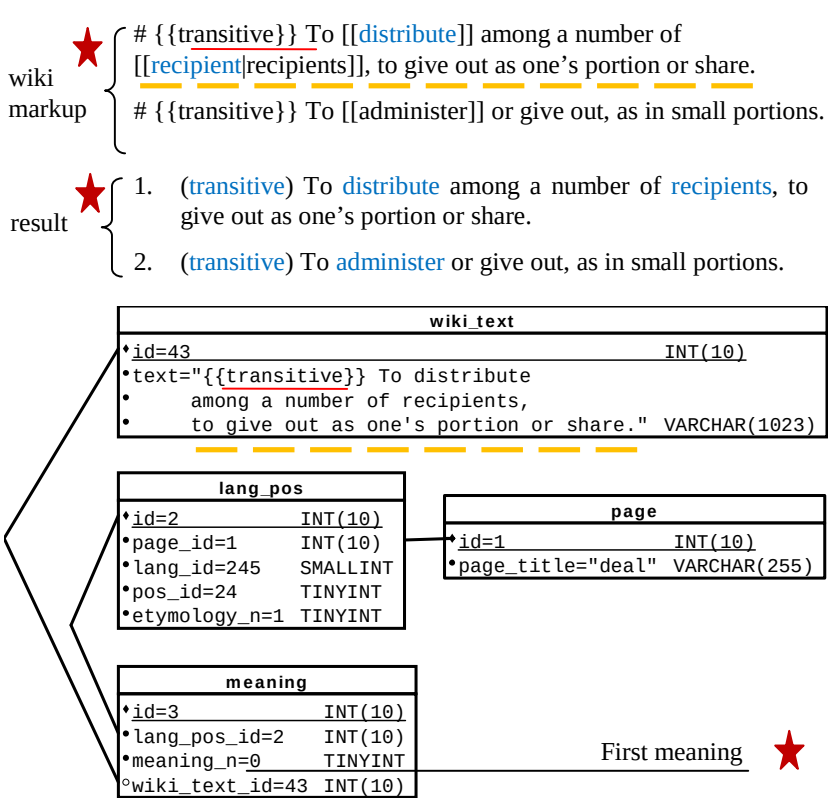

Рис. 2. Представление первого значения английского глагола «deal» в таблицах машинно-читаемого словаря

язык данной части словарной статьи (поле *lang_id*), часть речи (поле *pos_id*) и номер этимологии (поле *etymology_n*). Номер этимологии (пунктирный кружок на рис. 2) служит для различения омонимов. На рис. 1 указан порядковый номер второго омонима (*Etymology 2*), в базе данных значение поля *etymology_n* равно 1, нумерация идёт с нуля.

- *lang* – привязывает уникальный идентификатор к названию языка и двух или трёх буквенному коду языка (редко – более длинный код). В Английском Викисловаре с помощью бота построен полный список названий языков и их кодов, которые используются в этом словаре.[6] Парсер на данный момент распознаёт 169 языков из Английского Викисловаря[7] и 337 языков и их кодов в Русском Викисловаре.[8] На рис. 1 указан английский язык с кодом языка *en*, название языка и кода подчёркнуто двойной пунктирной линией.
- *part_of_speech* – содержит связку между названием части речи (поле *name*) и идентификатором (поле *id*), который используется в таблице *lang_pos* (поле *pos_id*). На рис. 1 указана часть речи (глагол, по англ. *verb*) подчёркнута пунктирной линией.

На рис. 1 первое значение, которое будет детально разбираться дальше, обозначено звёздочкой. На рис. 2 показан пример оформления значения слова в Английском Викисловаре: то, как это видит редактор словаря (*wiki markup*) и то, как это видит читатель словаря (*result*). Ниже на том же рис. 2 показаны таблицы машинно-читаемого словаря, ответственные за хранение данного значения словарной статьи.

Разработанный парсер извлекает из словарной статьи семантические отношения и переводы. Однако

[6] См. http://en.wiktionary.org/wiki/Wiktionary:Index_to_templates/languages

[7] См. http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:Translations

[8] См. http://ru.wiktionary.org/wiki/Участник:AKA_MBG/Статистика:Переводы

привязка переводов и семантических отношений идёт не ко всей статье (таблица *page*) и даже не к той её части, которая соответствует заданному языку и части речи (таблица *lang_pos*). В словарной статье, а, следовательно, и в машинно-читаемом словаре, привязка идёт к конкретному значению слова (например, на рис. 2 представлено первое значение, обозначенное звёздочкой, вики-текст которого подчёркнут пунктирной линией). Поэтому в машинно-читаемый словарь было необходимо добавить таблицу *meaning*, содержащую следующие поля:

- поле *lang_pos_id* указывает на часть речи, язык и название словарной статьи посредством таблицы *lang_pos*;
- поле *meaning_n* определяет номер значения слова (нумерация с нуля);
- поле *wiki_text_id* обеспечивает доступ к тексту толкования, который хранится в таблице *wiki_text* (подчёркнуто пунктирной линией).

Текст толкования (поле *text* таблицы *wiki_text* на рис. 2) на данный момент сохраняется почти целиком, как оно есть, по сравнению с исходным текстом в вики-разметке (см. *wiki_markup* вверху на рис. 2), т.е. без внутренних ссылок (квадратные скобки), но вместе со вспомогательной информацией в фигурных скобках (здесь шаблон *{{transitive}}*, указывающий на транзитивность данного значения глагола, подчёркнут сплошной линией). В будущем такие вспомогательные шаблоны будут распознаваться парсером, удаляться из текста толкования и сохраняться в отдельной таблице, привязанной к таблице *meaning*.

Некоторые поля таблиц базы данных могут принимать значение NULL. На рисунках это обозначено пустым кружком слева от названия поля таблицы. Например, на рис. 2 поле *wiki_text_id* в таблице *meaning* обозначено таким кружком.

Значение NULL в данном случае легко объяснимо. Поле *meaning.wiki_text_id* может не содержать ссылки на таблицу *wiki_text*, поскольку в словарной статье (например, в Русском Викисловаре) может отсутствовать толкование для одного из значений слова, однако могут быть указаны синонимы и переводы для этого отсутствующего толкования слова. При этом в машинно-читаемом словаре таблицы переводов (*translation* и *translation_entry*) и таблицы семантических отношений (*relation* и *relation_type*) привязываются к таблицам *lang_pos* и *meaning*, т.е. к конкретному значению, даже если текст толкования отсутствует.

Вики-ссылка или внутренняя ссылка (wikilink, internal link) – это ссылка, указывающая на другую статью в пределах данного вики-сайта. В случае Викисловаря, это обычно ссылка на нормализованную, основную форму слова. Например, вики-разметка на рис. 3 (вверху рисунка) содержит следующую внутреннюю ссылку: *[[recipient|recipients]]*. Текст в квадратных скобках, слева от вертикальной палочки – это основная форма слова, на которую ведёт ссылка (на рис. 3 подчёркнуто сплошной линией). Текст справа (на рис. 3 подчёркнуто пунктирной линией) – это текст, который будет виден читателю вики-страницы (см. *result* вверху на рис. 3).

При разработке парсера и структуры машинно-читаемого словаря было решено сохранять эту информацию (в таблицах *wiki_text_words*, *page_inflection*, *inflection*), поскольку с её помощью можно узнать:

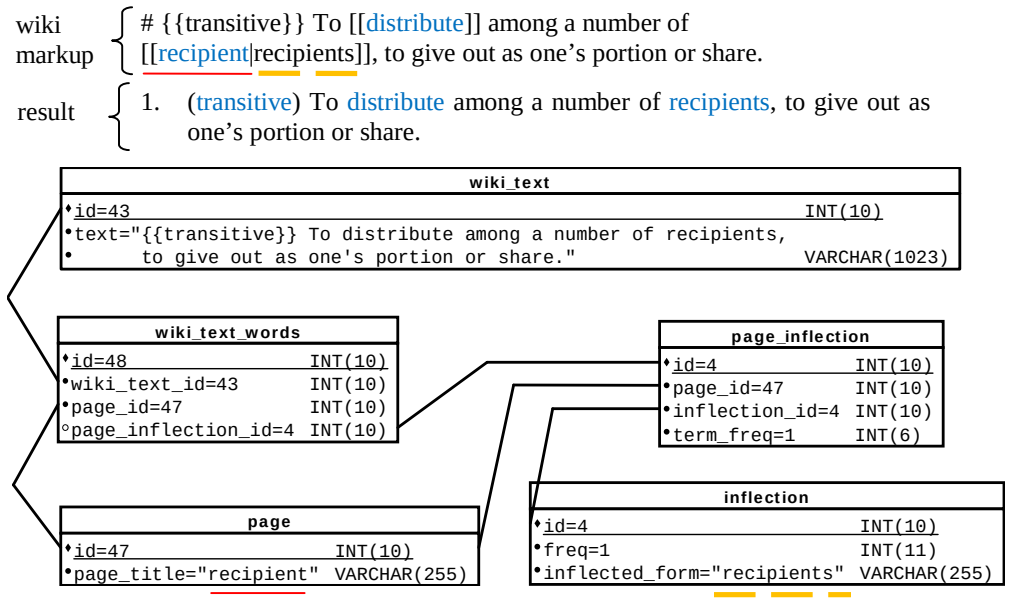

Рис. 3. Представление в таблицах *wiki_text_words*, *page_inflection*, *inflection* внутренних ссылок, обозначенных квадратными скобками ([[…]]) в вики тексте

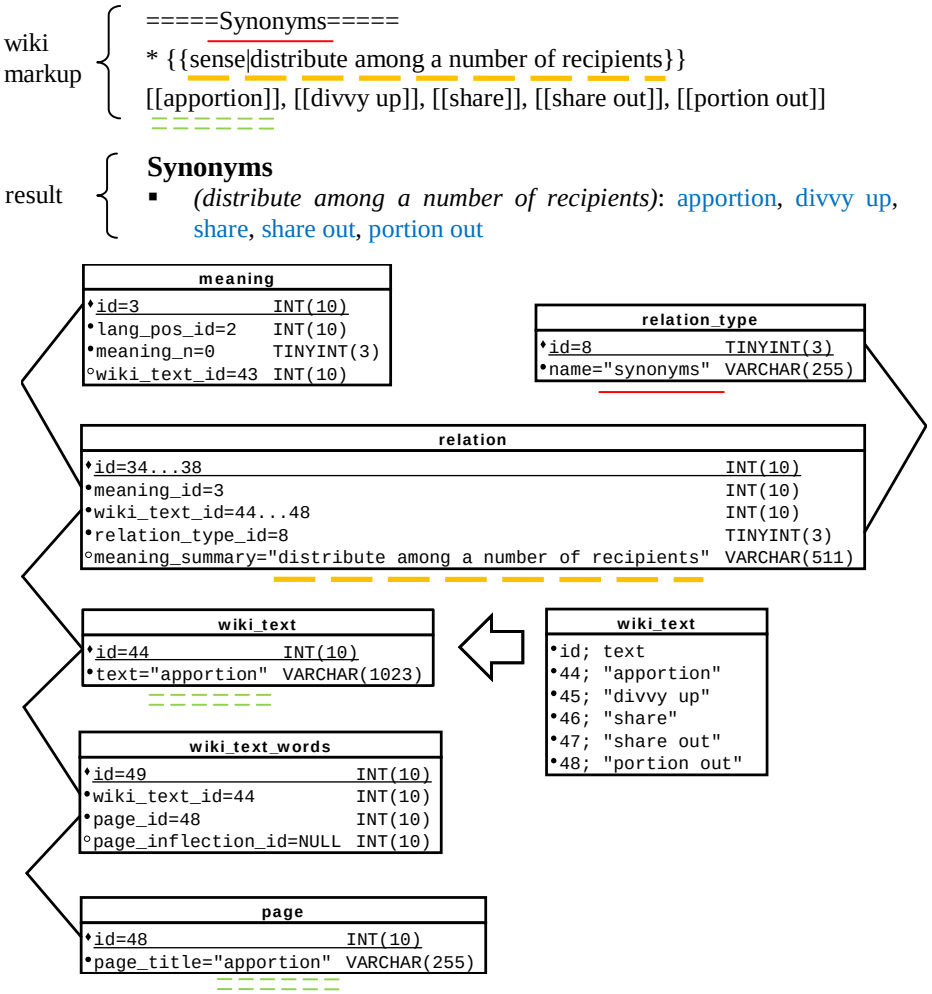

Рис. 4. Пример сохранения в БД списка синонимов (*apportion, divvy up, share, share out, portion out*) для первого значения (meaning_n=0) глагола *deal*

- какая форма для данного слова является основной? В идеале, информация о построении форм слова должна быть представлена в той словарной статье, куда ведёт ссылка. Теоретически Викисловарь должен описывать все слова, весь словарный запас языка. Конечно, эта цель ещё не достигнута. В Викисловаре часто встречаются внутренние ссылки красного цвета, указывающие цветом на то, что словарная статья для данного слова или фразы ещё не создана. Таким образом, для некоторых слов внутренние ссылки (на данный момент) могут быть единственным источником информации о формах слова.
- какие формы слова являются наиболее частотными для данного слова?
- какие слова являются ключевыми (или семантически близкими) для данного толкования? Словарные статьи, на которые ведут внутренние ссылки в толковании, в общем случае, могут не являться *ключевыми словами* толкования. Однако поскольку Викисловарь является ещё и тезаурусом, то эти слова могут использоваться для построения «среза» тезауруса, соответствующего данному толкованию. Этот «срез» может быть использован в различных задачах информационного поиска, в том числе для построения списка ключевых слов для данного толкования.

Для сохранения семантических отношений в машинно-читаемом словаре есть две таблицы: *relation* и *relation_type*. Таблица *relation_type* заполняется один раз (до работы парсера) и содержит девять типов семантических отношений, представленных в Английском Викисловаре.[9]

При извлечении семантических отношений из текста словарной статьи одновременно извлекается текст резюме толкования из шаблона *{{sense|}}* (на рис. 4

[9] См. http://en.wiktionary.org/wiki/Wiktionary:Semantic_relations

текст подчёркнут пунктиром) и сохраняется в поле *meaning_summary* таблицы *relation*.

На рис. 4 в таблице *relation* указаны интервалы значений в полях *id* и *wiki_text_id*. Это подчёркивает тот факт, что список синонимов (таблица *wiki_text*) привязан к данному значению слова (т.е. к таблице *meaning*) через таблицу *relation*.

Рис. 4 показывает, что при сохранении семантических отношений происходит дублирование информации в таблицах *wiki_text* и *page* (например, синоним *apportion* хранится в обеих таблицах, подчёркнут двойной пунктирной линией). Это частный случай, когда вики текст представляет собой одно слово, являющееся гиперссылкой (а именно: внутренней ссылкой) на какую-либо словарную статью. В более общем случае (см. предыдущий рис. 3) вики текст является фразой, которой не соответствует никакая словарная статья в Викисловаре, т.е. дублирования нет.

На рис. 5 показан пример оформления раздела «Переводы» в Английском Викисловаре: то, как это видит редактор словаря (*wiki markup*) и то, как это видит читатель словаря (*result*). На том же рисунке ниже в таблицах машинно-читаемого словаря указаны данные, связанные с переводом *dela* первого значения английского глагола *deal* на шведский язык. Перевод *dela* подчёркнут сплошной линией, язык – двойной пунктирной. Код шведского языка *sv*, информация о данном языке хранится в таблице *lang* в строке под номером 782. Поле *n_translation* в таблице *lang* указывает на число переводов с главного (в этом Викисловаре – английского) на заданный язык, т.е. в Английском Викисловаре содержится 16 тысяч переводов на шведский язык (по данным дампа Английского Викисловаря от 24 августа 2010 г.).

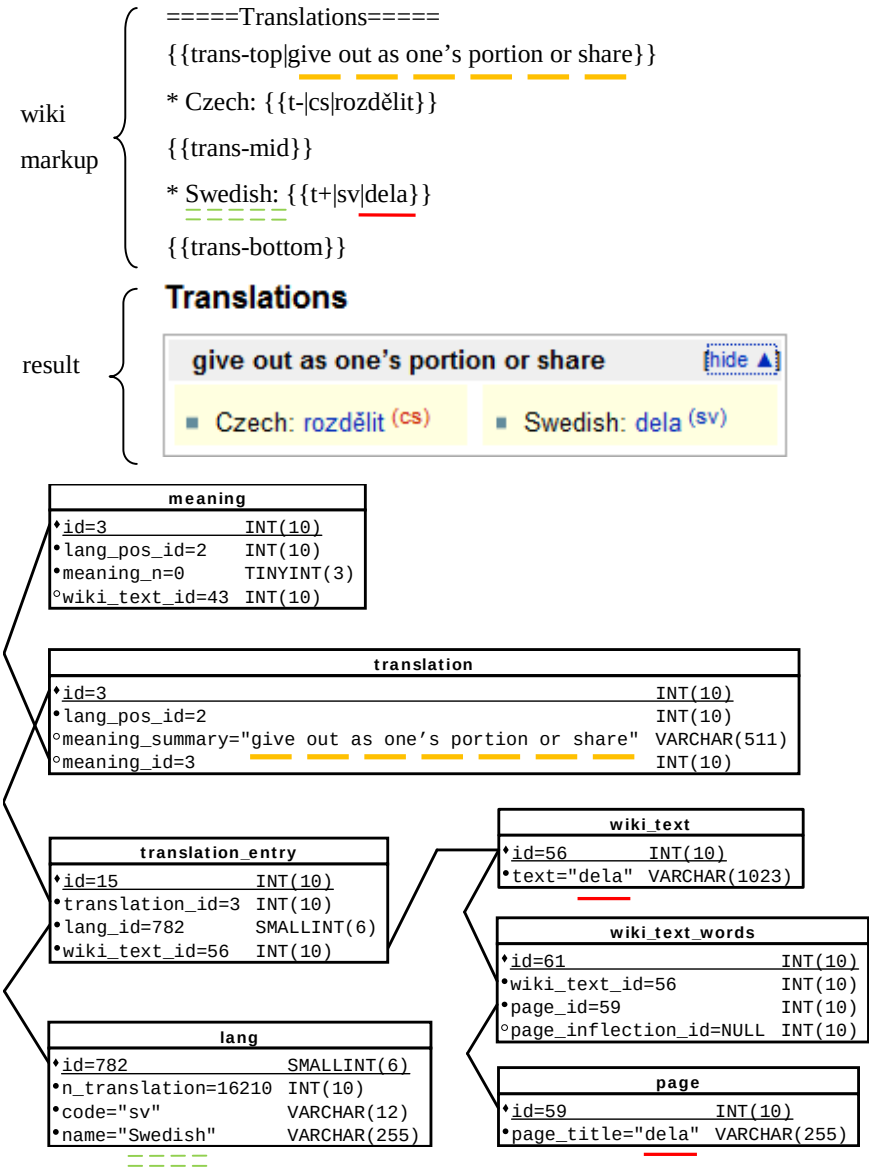

Рис. 5. Отображение данных раздела «Перевод» словарной статьи в таблицах машинно-читаемого словаря

Разделу «Перевод» соответствуют следующие связи между таблицами.

- Каждому значению слова (запись в таблице *meaning*) соответствует одна строка в таблице *translation* (т.е. отношение *один к одному*) Если для перевода указано краткое толкование (параметр шаблона *{{trans_top|}}*), то это толкование записывается в *translation.meaning_summary*, подчёркнуто на рисунке пунктирной линией.
- Одной записи в таблице *translation* может соответствовать множество записей в таблице *translation_entry*, по одной на каждый язык и на каждый перевод (отношение *один ко многим*). В примере на рис. 5 (вверху) указано два перевода, по одному на каждый язык, а именно: перевод на чешский (*rozdělit*) и на шведский (*dela*). Поэтому в таблице *translation_entry* будет ровно две записи со ссылками:
    - на текст перевода (поле *wiki_text_id*);
    - на язык перевода (поле *lang_id*);
    - на конкретное значение слова (*meaning.id*) посредством таблицы *translation* (поле *translation_id*).

Для полноты картины необходимо привести схему базы данных машинно-читаемого словаря на Рис. 7.

4. ЗАКЛЮЧЕНИЕ

Разработана архитектура модульного и расширяемого парсера Викисловаря. Реализованы модули для извлечения трёх блоков данных из словарных статей, а именно: значение слова, семантические отношения и перевод.

При анализе текста словарной статьи многократно используются разнообразные регулярные выражения, позволяющие вычленить из текста словарной статьи требуемую информации.[10] Такое вычленение и анализ страницы возможен только благодаря известной структуре статьи и применению шаблонов внутри статьи. И чем более жёсткая структура статьи принята сообществом данного Викисловаря, тем проще и надёжнее алгоритмы парсера. Чем больше шаблонов, широко используемых в Викисловаре, тем легче извлечь информацию, структурированную с их помощью.

В статье представлено преобразование «неявной» структуры, понятной только читателю словаря, в явную, «понятную» компьютерным программам форму. Т. е. показано соответствие между фрагментами (разделами) словарной статьи Викисловаря и таблицами базы данных машинно-читаемого словаря, в которые сохраняются данные, извлекаемые из Викисловаря. Такое автоматическое преобразование позволит легко использовать данные Викисловаря в различных проектах, связанных с обработкой текста.

Создание машинно-читаемых словарей является важным кирпичиков в основании небоскрёба автоматической обработки текста. MRD словари, и в том числе данные Википедии и Викисловарей, используются при построении онтологий [14], [17]; в машинном переводе [2], [11], при автоматическом упрощении текста [12], при поиске изображений [2], при разрешение лексической многозначности [9].

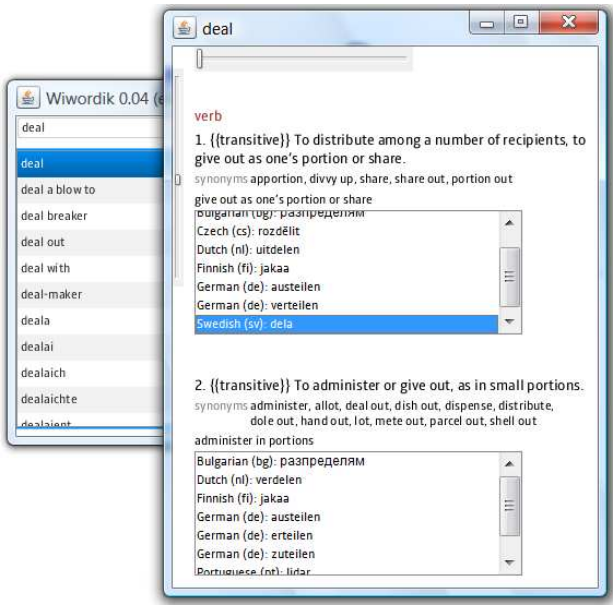

Рис. 6. Карточка словарной статьи “deal" является визуализацией данных машинно-читаемого словаря

Видно много заманчивых дорог, пойдя по которым можно развивать и парсер, и приложения на его основе. Однако в первую очередь нужно создать графический интерфейс к машинно-читаемому словарю, построенному по данным Английского Викисловаря (рис. 6). Для Русского Викисловаря такая оболочка уже создана и доступна онлайн.[11]

[10] См. http://ru.wiktionary.org/wiki/Викисловарь:Шаблоны

[11] См. программу *wiwordik* , построенную на основе данных Русского Викисловаря: http://code.google.com/p/wikokit/wiki/wiwordikRu

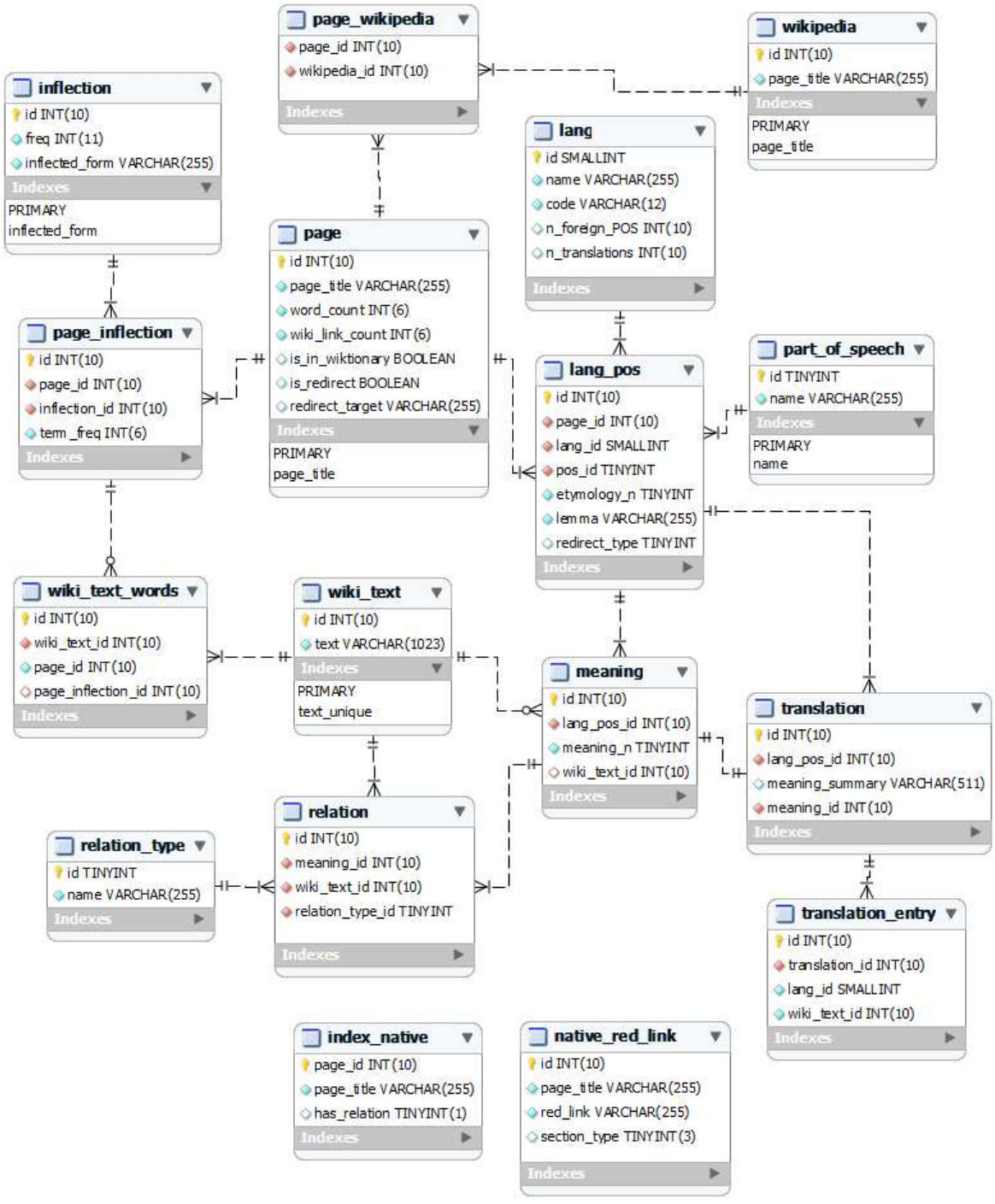

Рис. 7. Таблицы и отношения в базе данных машинно-читаемого словаря

ССЫЛКИ

[1] Фридл Дж. Регулярные выражения. Библиотека программиста. СПб.: Питер, 2001. – 352 с. ISBN 5-272-00331-4.

[2] Etzioni, O., Reiter, K., Soderland, S., and Sammer, M. Lexical Translation with Application to Image Search on the Web. In the proceedings of MT Summit XI. 2007. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.7536&rep=rep1&type=pdf

[3] Fiser D., Sagot B. Combining multiple resources to build reliable wordnets. In TSD 2008, Brno, Czech

Republic. 2008. http://alpage.inria.fr/~sagot/pub-en.html

[4] Herbelot A., Copestake A. Acquiring ontological relationships from wikipedia using rmrs. In Proceedings of the ISWC 2006 Workshop on Web Content Mining with Human Language Technologies. 2006. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.8469&rep=rep1&type=pdf

[5] Krizhanovsky A. A. The comparison of Wiktionary thesauri transformed into the machine-readable format. Language Resources and Evaluation. 2010. (submitted). http://arxiv.org/abs/1006.5040

[6] Krizhanovsky, A. A.; Feiyu Lin. Related terms search based on WordNet / Wiktionary and its application in Ontology Matching. In Proceedings of the 11th Russian Conference on Digital Libraries RCDL'2009. September 17-21, Petrozavodsk, Russia. 363-369. http://arxiv.org/abs/0907.2209

[7] Krizhanovsky, A. A. Index wiki database: design and experiments. In Proceedings of the Corpus Linguistics (The St. Petersburg, The Russia, October 6 - 10, 2008) CORPORA '08. 2008. http://arxiv.org/abs/0808.1753.

[8] Krizhanovsky, A. A. Synonym search in Wikipedia: Synarcher. In Proceedings of the 11-th International Conference "Speech and Computer" (The St. Petersburg, Russia, June 26 - 29, 2006). SPECOM '06. 474-477. 2006. http://arxiv.org/abs/cs/0606097

[9] Krovetz R., Croft W. B. Word sense disambiguation using machine-readable dictionaries. In Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval, p.127-136, June 25-28, 1989, Cambridge, Massachusetts, United States. http://elvis.slis.indiana.edu/irpub/SIGIR/1989/pdf14.pdf

[10] Meyer C. M., Gurevych I. Worth its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet. In Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics, p. 38-49. Iasi, Romania, 2010. http://www.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2010/cicling2010-meyer-lsrcomparison.pdf

[11] Muller, C., Gurevych, I. Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark, September 17-19. 2008. http://www.clef-campaign.org/2008/working_notes/mueller-paperCLEF2008.pdf

[12] Napoles C., Dredze M. Learning Simple Wikipedia: A Cogitation in Ascertaining Abecedarian Language. Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids at NAACL-HLT. 2010. http://www.cs.jhu.edu/~mdredze/

[13] Sumida A., Torisawa K. Hacking Wikipedia for Hyponymy Relation Acquisition. In Proceedings of International Joint Conference on NLP (IJCNLP'08). 2008. http://acl.eldoc.ub.rug.nl/mirror/I/I08/I08-2126.pdf

[14] Wandmacher, T., Ovchinnikova, E., Krumnack, U. and Dittmann, H. Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. In Proc. Third Australasian Ontology Workshop (AOW 2007), Gold Coast, Australia. CRPIT, 85. Meyer, T. and Nayak, A. C., Eds. ACS. 61-69. 2007. http://crpit.com/abstracts/CRPITV85Wandmacher.html

[15] Wilms G. J. Computerizing a Machine Readable Dictionary. In Proceedings of the 28th annual Southeast regional conference, ACM Press. 306–313. 1990. http://computerscience.uu.edu/faculty/jwilms/papers/acm90/acm90.pdf

[16] Zesch T., Mueller C., Gurevych I. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In Proceedings of the Conference on Language Resources and Evaluation (LREC). 2008. http://elara.tk.informatik.tu-darmstadt.de/publications/2008/lrec08_camera_ready.pdf

[17] Рубашкин, В.Ш.; Бочаров, В.В.; Пивоварова, Л.М.; Чуприн Б.Ю. Опыт автоматизированного пополнения онтологий с использованием машиночитаемых словарей. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). - М.: Изд-во РГГУ, 2010. http://www.dialog-21.ru/dialog2010/materials/html/63.htm