ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004.65:004.4’12:809.51

О.А. Резник, О.Н. Ванеев

РАЗРАБОТКА КИТАЙСКО-РУССКОГО СЛОВАРЯ НА ОСНОВЕ ОТОБРАЖЕНИЕ ЕГО КОНЦЕПТУАЛЬНОЙ СТРУКТУРЫ В ВИДЕ РЕЛЯЦИОННОЙ БАЗЫ ДАННЫХ

Введение

В настоящее время все чаще возникает необходимость в осуществлении переводов различных документов как с китайского языка на русский, так и с русского на китайский . А значит, растет и потребность в программных средствах, помогающих осуществлять перевод быстрее и качественнее.

В данной работе предлагается подход к созданию электронного китайско-русского словаря на основе отображения состава базовых элементов китайского языка и связей между ними в виде реляционных отношений. Предполагается, что данный подход позволит не только создать средства автоматизации перевода, но и позволит отобразить концептуальную структуру языка.

Китайский язык обладает рядом особенностей, которые необходимо учитывать при создании электронного словаря [1].

1. Специфика китайского языка.

В китайском языке слова записываются не буквами, а **иероглифами**, что образует множество трудностей при создании словарей.

Иероглифы состоят из строго определенных базовых графических элементов - **черт**, всего их 24. В каждом иероглифе можно посчитать количество черт. Порядок написания черт в иероглифе задан правилами каллиграфии.

Также иероглиф может быть разложен на составные части, называемые **ключами**. Некоторые ключи являются самостоятельными иероглифами и имеют собственное значение, а некоторые - встречаются только в сочетании с другими, лишь в составе более сложных иероглифов.

Например, иероглиф 星 xīng (звезда) состоит из двух ключей, являющихся самостоятельными иероглифами: 日 rì (день, солнце) и 生 shēng (рождаться, сырой)

Фонетика иероглифа – это его произношение. Ее определяют **слог** и **тон**.

В китайском языке 422 **слога**. Транскрипция слогов записывается латинскими буквами (например: mang, pa, bu). Множество иероглифов, зачастую имеющих почти противоположное значение, произносятся одним и тем же слогом. Это очень затрудняет поиск иероглифов в словаре.

Тон определяет интонацию произношения. В китайском языке 5 тонов:

- нулевой (обозначается o) – нет интонации
- первый (обозначается ‾) – ровная интонация
- второй (обозначается ╱) – интонация вверх
- третий (обозначается ∨) – интонация сначала вниз, а потом вверх
- четвертый (обозначается ╲) – интонация вниз

Один и тот же слог, произнесенный разными тонами, может иметь абсолютно разные значения. Например, слог ma произнесенный первым тоном означает «мама», и этот же слог, произнесенный 3-м тоном, означает «лошадь». Уникальные сочетания слога и тона рассматриваются как разные фонетики.

Фонетика иероглифа традиционно обозначается как слог с нарисованным над ним обозначением тона.

Каждому иероглифу соответствует как минимум одна фонетика. Но чаще один и тот же иероглиф может иметь несколько фонетик, что соответствует разным его значениям. В таком случае его значение выбирается по контексту.

В большинстве случаев иероглиф (с одной из его фонетик) имеет самостоятельное значение и соответствует одному слову.

Также иероглиф может быть частью **сочетания**, такое сочетание может означать одно слово или целую фразу. При этом иероглифы, входящие в его состав, по отдельности имеют значение, не связанное со значением всего сочетания.

Например, иероглиф 星 xīng переводится как «звезда» или «искра», а также является частью многих сочетаний, например: 星期 xīngqī, что означает «неделя».

Таким образом, под одним словом в китайском языке будем подразумевать иероглиф с од-

ной из его фонетик, имеющий собственное значение, или сочетание из нескольких таких иероглифов.

2. Определение функциональных требований к электронному словарю.

Чтобы сделать использование словаря максимально удобным, целесообразно обеспечить 3 основные функции:

- вывода информации об внесённых в него иероглифах;
- вывода информации о сочетаниях иероглифов;
- вывода перевода с русского на китайский язык;

При переводе **с китайского языка на русский** ввод иероглифов может осуществляться: фонетически (ввод слогов латинскими буквами), с помощью ключей (если неизвестно, как иероглиф читается), по количеству черт и непосредственно ввод иероглифа.

Фонетический ввод иероглифов принято осуществлять таким образом:

1) пользователь вводит слог с клавиатуры;
2) на экран выводится список иероглифов, имеющих такую фонетику;
3) пользователь выбирает из списка необходимый иероглиф.

Ввод с помощью ключей:

1) пользователь вводит количество черт, из которых состоит ключ, являющийся частью искомого иероглифа;
2) на экран выводится список ключей с введенным количеством черт;
3) пользователь выбирает ключ;
4) на экран выводится список иероглифов, в состав которых входит выбранный ключ;
5) если таких иероглифов слишком много, пользователь вводит количество черт, которые нужно добавить к данному ключу, чтобы получить вводимый иероглиф;
6) на экран выводится список иероглифов, содержащих выбранный ключ и введенное количество черт;
7) пользователь выбирает иероглиф из списка;

Ввод по количеству черт:

1) пользователь вводит количество черт в иероглифе;
2) на экран выводится список иероглифов с введенным количеством черт;
3) пользователь выбирает иероглиф;

Информация об иероглифе должна содержать:

- иероглиф;
- его фонетику: слог и тон (если существует несколько вариантов фонетики данного иероглифа, должна быть обеспечена возможность просмотра перевода иероглифа с другой фонетикой);
- перевод иероглифа с данной фонетикой (если иероглиф имеет несколько значений, то должны быть указаны они все);
- сочетания, в которых участвует данный иероглиф с переводом и фонетикой всех иероглифов этого сочетания;

Так как некоторые значения употребляются чаще других, желательно выводить значения в порядке убывания частоты их употребления (от наиболее к наименее часто употребляемым).

При переводе **с русского на китайский язык**, слова должны вводиться с клавиатуры.

При этом на экран необходимо вывести:

- все иероглифы и сочетания, имеющие данное значение;
- их фонетику (слоги и тоны).

Необходимо создание двух клиентских приложений:

- клиент для редактирования содержания словаря
- клиент для просмотра данных, содержащихся в словаре

3. Реализация словаря на основе реляционной базы данных

В настоящее время на кафедре ИиАПС КуЗГТУ реализован прототип информационной системы. Создана база данных и прототипы клиентских приложений.

Выделенные отношения и связи между ними.

При разработке модели базы данных использовались положения, изложенные в [2]

Так как один иероглиф может иметь различную фонетику, для избежания увеличения объема хранимых данных, было решено выделить отдельное отношение для хранения иероглифов.

В базе данных это отношение представлено таблицей **«Иероглифы» (ieroglifi).** Она содержит следующие поля:

- номер иероглифа (id_ier) – уникальный номер для каждого иероглифа, с помощью него будет осуществляться доступ к иероглифу, это первичный ключ;
- иероглиф (ier) – символ иероглифа в формате Unicode;
- количество черт (kol_chert) – количество черт в иероглифе, необходимо для осуществления ввода иероглифа по количеству черт.

Для хранения всех возможных вариантов слогов (в китайском языке всего 422 слога) используется таблица **«Слоги» (slogi)**, ее поля:

- номер слога(id_slog) – уникальный номер, с помощью него осуществляется доступ к записи слога, это первичный ключ для данного отношения.
- слог (slog) – запись слога, которым читается иероглиф, латинскими буквами.

Для хранения данных о возможных вариантах фонетики иероглифов было сформировано отношение **Фонетика» (fonetika)**, оно содержит следующие атрибуты:

- номер иероглифа с фонетикой (id_iersfon) – уникальный номер, с помощью него осуществляется доступ к значениям иероглифа и сочетаниям, в которые входит иероглиф, это первичный ключ для данного отношения;
- номер иероглифа (id_ier) – ключ, связывающий символ иероглифа с его фонетикой;
- номер слога (id_slog) - ключ, связывающий иероглиф и слог, которым он читается;
- тон (ton)– число от 0 до 4, обозначающее тон иероглифа.

Так как один иероглиф может иметь несколько фонетик, связь между отношениями «иероглифы» и «фонетика» является связью «один ко многим» (1-М).

Для осуществления ввода иероглифов с помощью ключей, были введены отношения «ключи» и «разложение по ключам».

Таблица «Ключи» (kluchi) содержит поля:

- номер ключа (id_kl) – уникальный номер для каждого ключа, с помощью него осуществляется доступ к изображению ключа, это первичный ключ для данного отношения;
- рисунок (risunok) – изображение ключа;
- название ключа (nazv);
- количество черт в ключе (kol_chert).

Поскольку один ключ может содержаться во многих иероглифах, и иероглиф может состоять из нескольких ключей, то связь между отношениями «ключи» и «иероглифы» является связью «многие ко многим» (М - М). Эта связь задается вспомогательным отношением **«Разложение по ключам» (razl_kluch_ier)**, атрибуты которого:

- номер строки (id) – уникальный номер для каждой строки в таблице, первичный ключ;
- номер иероглифа (id_ier) – первичный ключ таблицы «иероглифы»;
- номер ключа (id_kluch) – первичный ключ таблицы «ключи».

Таблица **«Сочетания» (soch)** хранит количества иероглифов в сочетаниях и содержит поля:

- номер сочетания (id_soch) – номер, уникальный для каждого сочетания, с помощью него осуществляется доступ к сочетанию, это первичный ключ данной таблицы;
- количество иероглифов в сочетании (col).

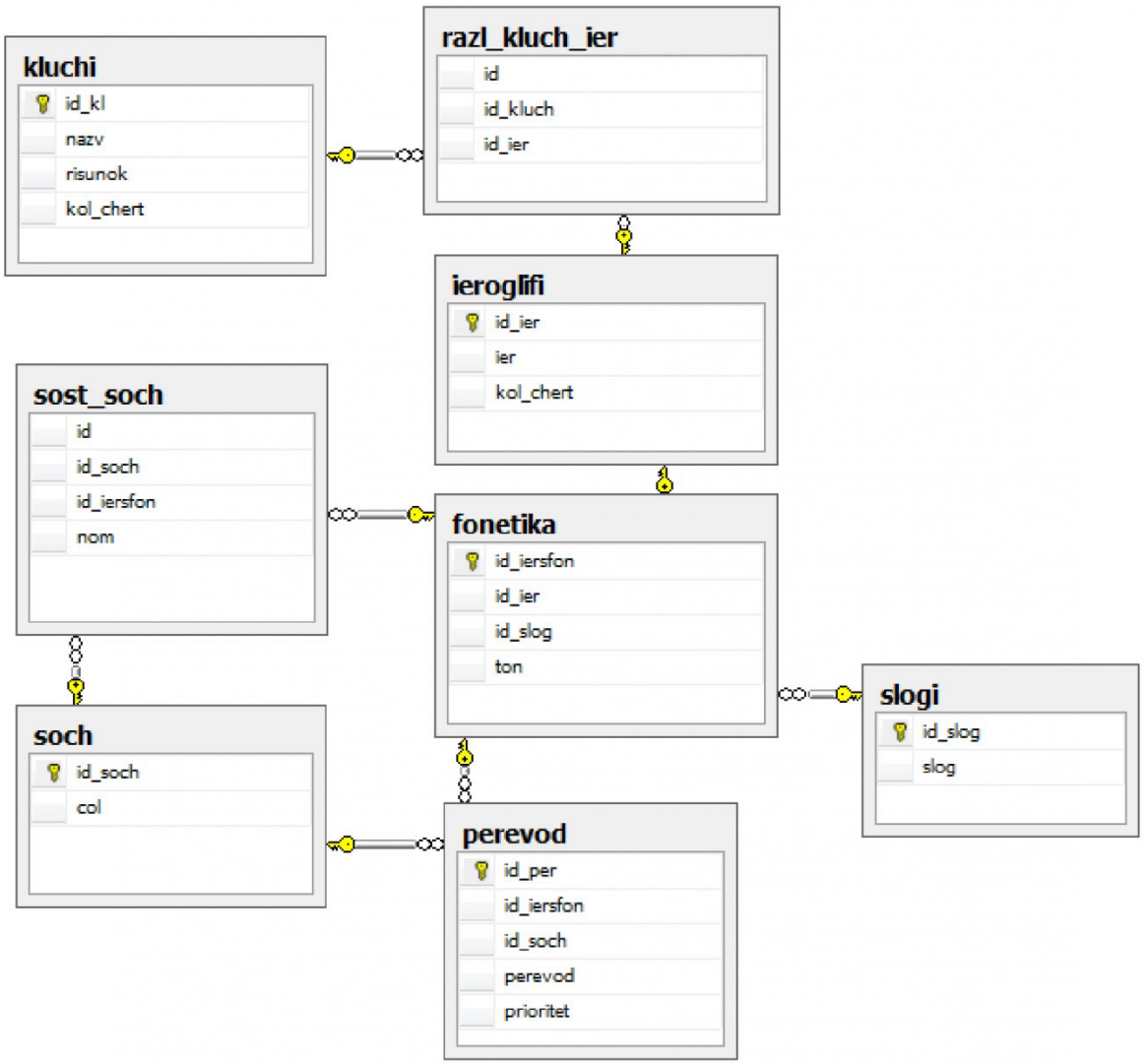

Рис.1 Концептуальная модель базы данных.

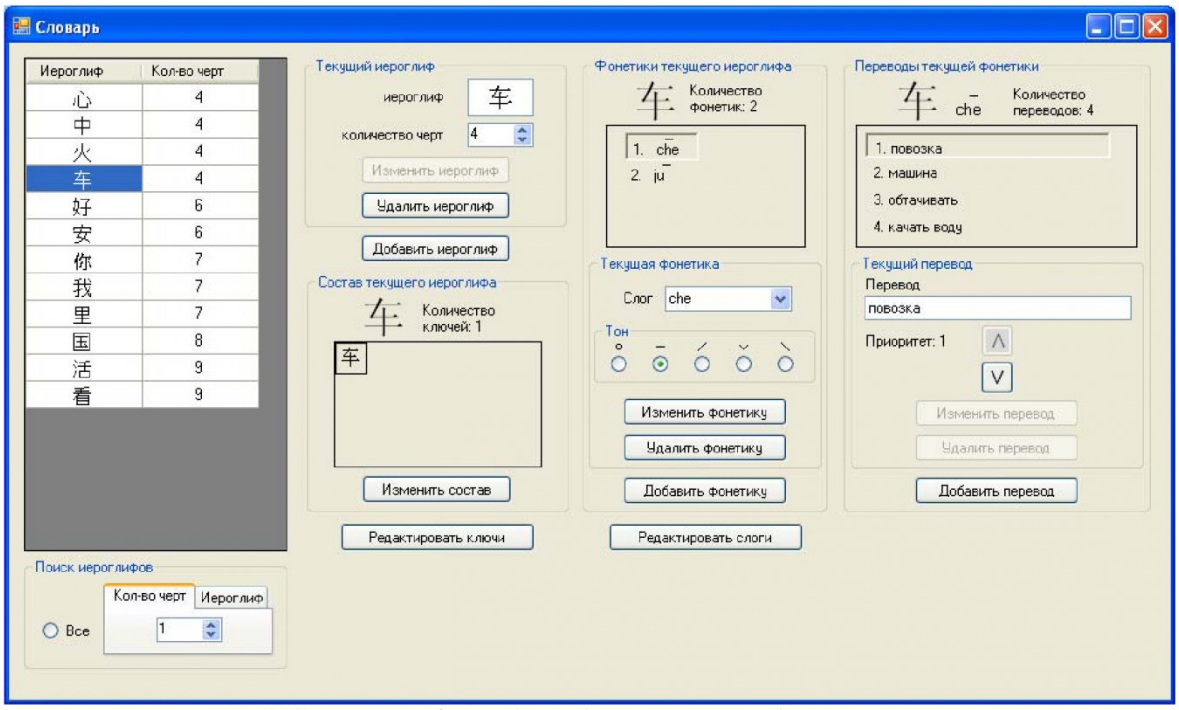

Рис. 2 Клиентская форма для редактирования содержания словаря в рабочем состоянии

Для хранения состава сочетаний используется отношение **«Состав сочетаний» (sost_soch),** которое содержит следующие поля:

- номер строки (id) – уникальный номер для каждой строки в таблице, первичный ключ;
- номер сочетания (id_soch) – номер,

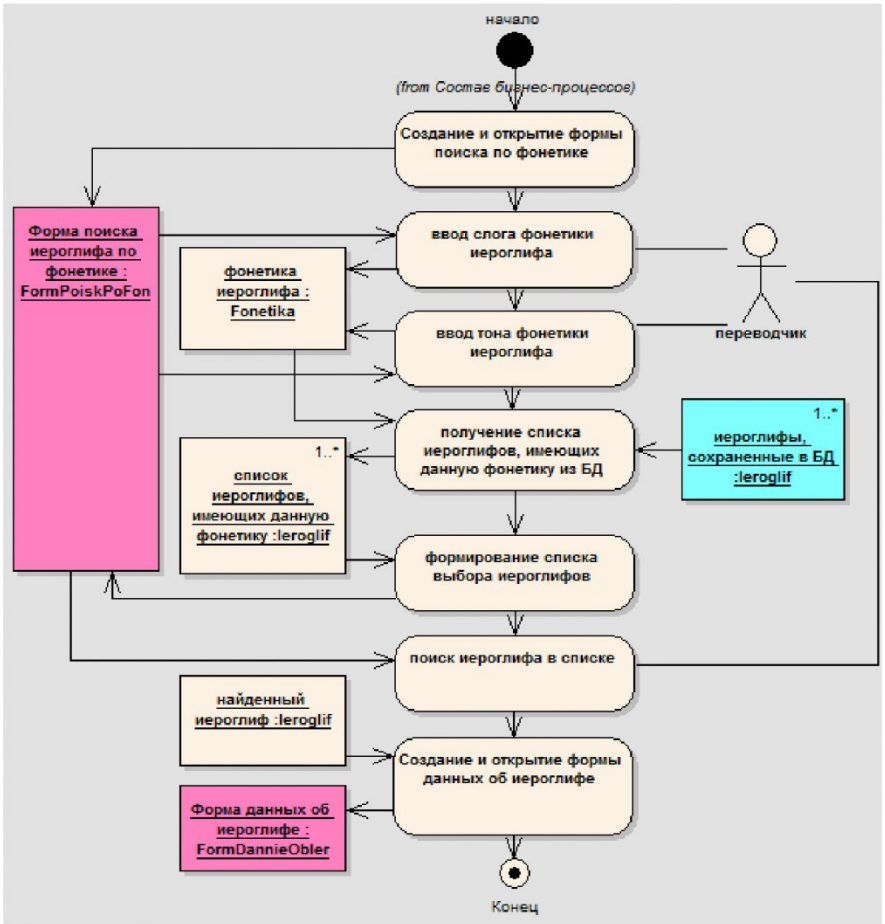

Рис. 3. Сценарий «Ввод иероглифа по фонетике»

уникальный для каждого сочетания, с помощью него осуществляется доступ к сочетанию;

- номер иероглифа с фонетикой, входящего в сочетание (id_iersfon);
- порядковый номер в сочетании (nom) – место, занимаемое данным иероглифом в сочетании.

Связь между отношениями «Сочетания» и «Состав сочетаний» является связью «один ко многим» (1-М), так как в одном сочетании содержится несколько иероглифов с фонетикой.

Поскольку один иероглиф с определенной фонетикой может входить в состав нескольких сочетаний, связь отношений «Фонетика» и «Состав сочетаний» является связью «многие к одному» (М-1)

Перевод иероглифов с определенной фонетикой и сочетаний хранится в отношении **«Перевод» (perevod)**. Атрибутами данного отношения являются:

- № перевода (id_per)– первичный ключ данного отношения;
- номер иероглифа с фонетикой (id_iersfon) – первичный ключ отношения «Фонетика» с помощью него осуществляется доступ к иероглифу с фонетикой;
- номер сочетания (id_soch) – первичный ключ отношения «сочетания» с помощью него осуществляется доступ к сочетанию;
- значение перевода (perevod);
- приоритет значения(prioritet) – число, указывающее на то, как часто употребляется данный иероглиф с таким значением. Чем меньше число, тем чаще употребляется данное значение.

В данной таблице поля «номер иероглифа с фонетикой» и «номер сочетания» взаимно исключают друг друга. То есть, если задан номер иероглифа, то номер сочетания должен не содержать значение, и наоборот.

В связи с тем, что один иероглиф с фонетикой или сочетание могут иметь несколько значений, то связи между отношениями «Фонетика»- «Перевод» и «Сочетания»-«Перевод» являются связями «один ко многим» (1-М)

Диграмма «сущность- связь» отображающая содержание разработанной модели базы данных представлена на рис. 1.

Прототипы клиентских приложений

Поскольку разработка клиентских приложений еще не закончена, реализованы не все необходимые функции. Описание созданных прототипов приведено ниже.

1. Клиент для редактирования содержания словаря.

На данный момент реализованы функции:

- поиска иероглифа по количеству черт и по иероглифу;
- добавления, удаления и изменения иероглифа;
- добавления, удаления и изменения ключа;
- отображение и редактирование состава иероглифа;
- добавления, удаления и изменения слога;

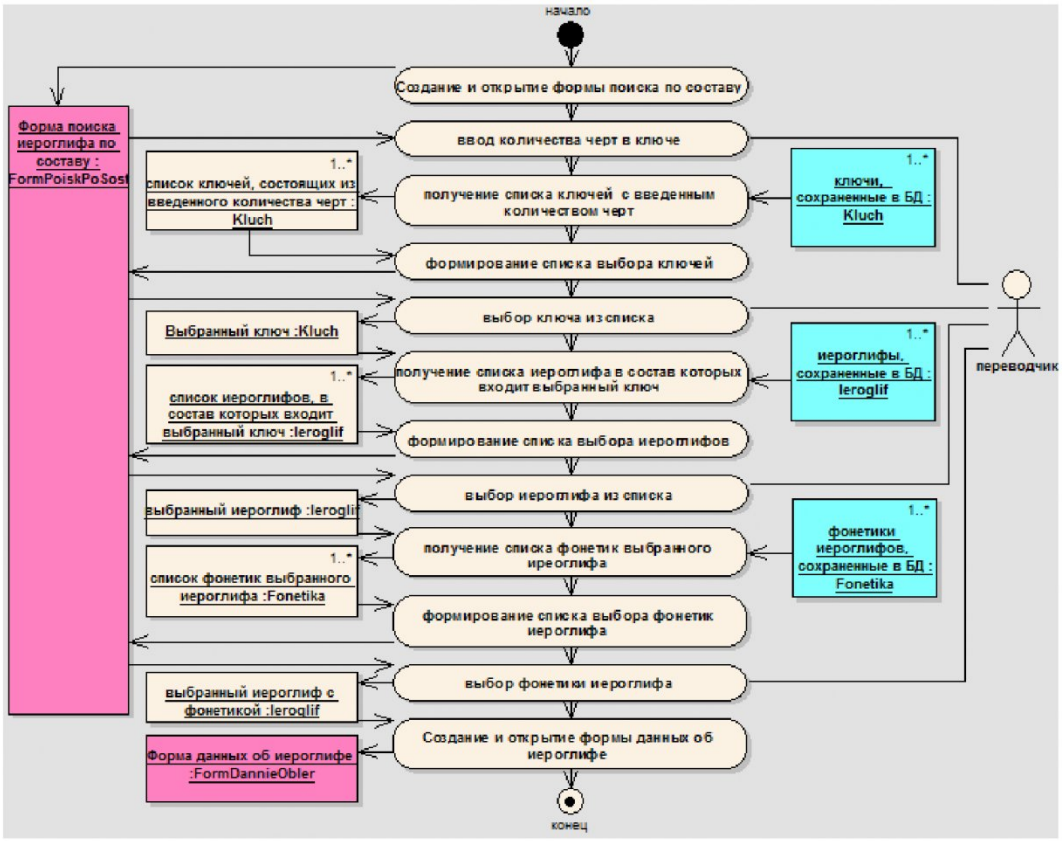

Рис.4 Сценарий «Ввод иероглифа по ключам»

- отображения и редактирования фонетик иероглифа;
- отображения и редактирования переводов фонетики

Пока не реализованы функции:

- отображения и редактирования сочетаний и их состава;
- отображения и редактирования переводов сочетаний/

Пример клиентского приложения для редактирования содержания словаря в рабочем состоянии представлен на рис. 2.

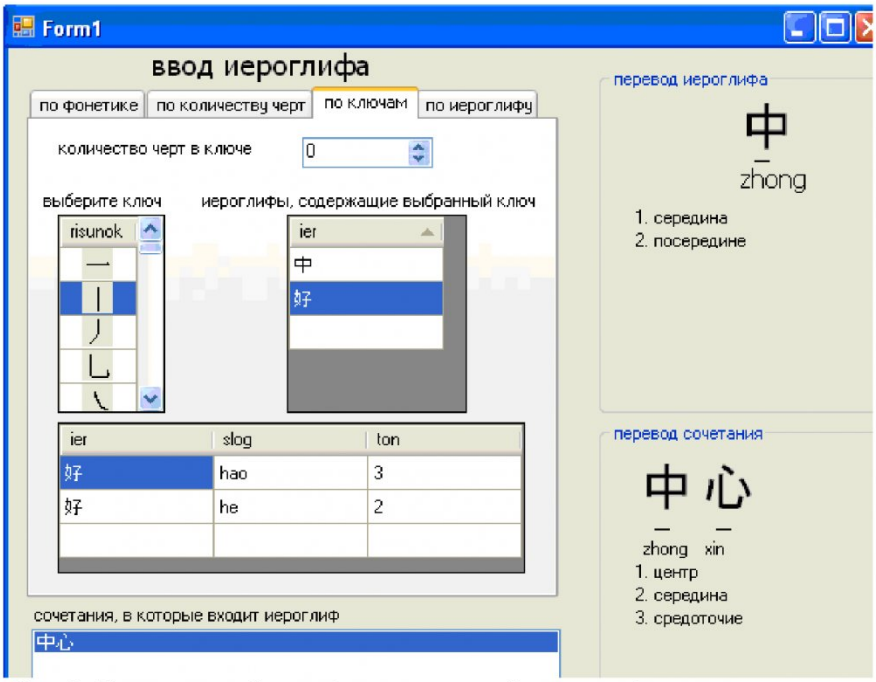

Рис. 5 Клиентская форма для просмотра данных, содержащихся в словаре в рабочем состоянии

2. Клиент для просмотра содержания словаря.

В созданном прототипе клиентского приложения реализованы функции:

- ввод иероглифа по фонетике, по ключам, по количеству черт и ввод иероглифа;
- отображение перевода иероглифа с фонетикой;
- отображение сочетаний, в состав которых входит иероглиф;
- отображение перевода сочетания.

Пока не реализована функция перевода с русского на китайский язык.

Содержание сценария «Ввод иероглифа по фонетике» представлено в виде диаграммы на рис. 3.

Содержание сценария «Ввод иероглифа по ключам» представлено диаграммой на рис. 4.

Пример клиентского приложения для просмотра содержания словаря в рабочем состоянии представлен на рис. 5.

Заключение

Таким образом, в настоящее время реализован начальный вариант информационной системы позволяющий выполнять основные функции редактировании внесенных данных и выполнения перевода отдельных иероглифов и сочетаний

Предполагается при дальнейшей разработке обеспечить более удобный ввод информация об иероглифах, сочетаниях и их переводе.

Необходимо повышать функционал системы, как реализуя описанные выше функции, так и добавляя следующие возможности:

- возможность разбиения слов на смысловые группы (например: города, цвета, имена)
- возможность задания частей речи
- возможность добавления примеров предложений
- возможность распечатки информации об иероглифах, сочетаниях и их переводе

Для реализации данных функций, необходимо развивать саму концептуальную модель, на основе которой строится база данных и словарь в целом.

СПИСОК ЛИТЕРАТУРЫ

1. *Горелов В. И.* Теоретическая грамматика китайского языка: Учебное пособие для студентов пед. ив-тов по спец. Иностр. яз.— М.: Просвещение, 1989

2. *Роб П., Коронер Л.* Системы баз данных: проектирование, реализация, управление. – 5-е изд., перераб. и доп.: Пер. с англ. – СПб.: БХВ-Петербург, 2004. – 1040 с.: ил.

□ Авторы статьи :

Резник
Ольга Анатольевна
- студентка гр. ИТ-042
КузГТУ
Тел. 9059141477

Ванеев
Олег Николаевич
-канд.техн.наук, доц.каф.
информационных и автоматизированных производственных систем КузГТУ
Тел. 923-611-797