УДК 519.7:007.52; 519.711.3

И.Д. Вечирская

Харьковский национальный университет радиоэлектроники,
г. Харьков, Украина, ira_se@list.ru

РАЗРАБОТКА ТРЕХЯЗЫЧНОГО ТЕРМИНОЛОГИЧЕСКОГО СЛОВАРЯ НА ОСНОВЕ АЛГЕБРЫ КОНЕЧНЫХ ПРЕДИКАТОВ

Статья посвящена разработке лексикографической базы данных трехязычного терминологического словаря. Проведен детальный анализ схемы связей между таблицами и структуры самих таблиц с помощью метода трехслойной декомпозиции предиката, что позволило определить пути решения проблемы создания прямых и обратимых трехязычных электронных словарей.

АЛГЕБРА КОНЕЧНЫХ ПРЕДИКАТОВ, ЛЕКСИКОГРАФИЧЕСКАЯ БАЗА ДАННЫХ, МЕТОД ТЕРХУРОВНЕВОЙ ДЕКОМПОЗИЦИИ ПРЕДИКАТОВ, ТЕРМИНОЛОГИЧЕСКИЙ СЛОВАРЬ, ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ

Введение

Существует несколько подходов к построению словарей. Один из них, теоретически правильный и обоснованный, предполагает детальную проработку всех методологических аспектов, построение концептуальной модели лексикографической системы [1] и последующей микроструктуры частей словарных статей. Таким образом, на каждом шаге происходит все более углубленное исследование предметной области, ее своеобразная детализация. Возможен также другой путь исследований, где производится переход от частного к общему. В таком случае уже в самом начале производится подробное изучение принципов организации данных, исследуются всевозможные связи с целью их дальнейшего обобщения и получения закономерностей. И только потом осуществляется построение модели системы.

На практике же редко получается отыскать решение задачи тем или иным способом и даже четко определить подход к поиску решения. Как правило, нахождение решения – это перебор или объединение нескольких методов решения с периодической сменой самих подходов. Так, при построении трехязычного терминологического словаря сначала стояла задача непосредственного составления самого словаря в электронном виде, часть которого уже была представлена документом Word, с целью дальнейшего вывода на печать. Реестр словаря составляют однословные термины и терминологические словосочетания по информатике и радиоэлектронике. Кроме терминов в реестр словаря включены некоторые общелитературные слова и словосочетания для облегчения чтения и составления текстов на украинском и английском языках.

1. Построение лексикографической базы данных терминологического словаря

Создание электронных словарей включает, как правило, следующие этапы обработки [1]:

1) Путем сканирования и распознавания получают электронный вариант текста;

2) Электронный текст словаря представляется в виде массива отдельных словарных статей;

3) По формальным признакам автоматически проводится декомпозиция массива словарных статей.

Принцип построения терминологического словаря – алфавитно-гнездовой [2]. Заголовочным словом является русское слово-термин. Гнездо включает терминологические словосочетания, элементом которых является заголовочный термин. Внутри гнезда терминологические словосочетания располагаются в алфавитном порядке. Если в заголовочную часть входит несколько однословных терминов (термины-синонимы или видовые пары глаголов), то для глаголов первым идет глагол несовершенного вида, для других частей речи – наиболее общеупотребительный. Часть заглавного слова, общая для всех терминологических словосочетаний в гнезде, отделяется прямой жирной чертой (|), а в производных словах вместо неё выступает тильда (~). Если общая часть не выделена, то вместо тильды подставляется все слово.

Видовые пары глаголов разделяются косой чертой, первым идёт глагол несовершенного вида. Все русские термины и терминологичекие словосочетания приведены жирным шрифтом.

Терминологические словосочетания строятся таким образом, чтобы тильда была на первом месте. Соответствующие украинские переводные эквиваленты сохраняют порядок слов русского словосочетания, английские эквиваленты сохраняют порядок слов, естественный для текстов.

Примеры формальных (полиграфических) признаков приведены в табл.1.

Схема связей между таблицами лексикографической базы данных трехязычного терминологического словаря представлена на рис. 1.

После автоматического построения лексикографической база данных была ее необходимо откорректировать.

Таблица 1

Структурный элемент	Формальный признак
Слово	Начало: абзац, буква (украинский, полужирный шрифт). Окончание: запятая; точка; пробел, индекс, цифра, запятая; пробел *див*; пробел, индекс, цифра, пробел *див*; двоеточие.
Стилистические и грамматические ремарки	Начало: текст после цифры (курсив). Окончание: текст (обычный шрифт).

2. Технологические аспекты разработки баз данных словаря и выбор программных средств

При разработке русско-украинско-английского словаря по информатике и радиоэлектронике использовалась СУБД Firebird. На ее выбор оказали влияние следующие факторы: соответствие SQL стандартам; поддержка транзакции, хранимых процедур; наличие утилиты бэкапа; возможность легкого переноса с одного компьютера на другой; наличие модуля для работы с БД без установки сервера Firebird, что упрощает его использование и инсталляцию на машинах пользователей; поддержка многопользовательского режима; а также наличие поддержки в основных системах разработки: PHP, Delphi, Perl, Borland C++; быстрое создание баз данных и работа с ними.

Инструментарий пользователя для заполнения словаря разрабатывался в среде Delhi 7. Веб-интерфейс не был задействован, т.к. он менее удобен в плане интерактивности, что важно для оператора, который будет заполнять словарь. Планируется использовать веб-интерфейс для предоставления словаря широкому кругу пользователей: поиск, вывод результатов, что предполагает задействование синхронизации данных.

В настоящее время словарь находится в стадии разработки. Изначально словарь был предназначен для ввода словарных статей, хранения данных, удобного поиска с возможностью редактирования и экспорта в Word с последующей печатью словаря. На данном этапе в базу данных входит 15 таблиц (рис. 1), которые отображают связи между словом-термином, его терминологическими словосочетаниями на русском языке и соответствующими переводными эквивалентами на украинском и английском языках. Грамматические показатели отображают следующие признаки частей речи:

– для существительного – род (*ж.*, *с.*, *ч.*,), число (*мн.*);

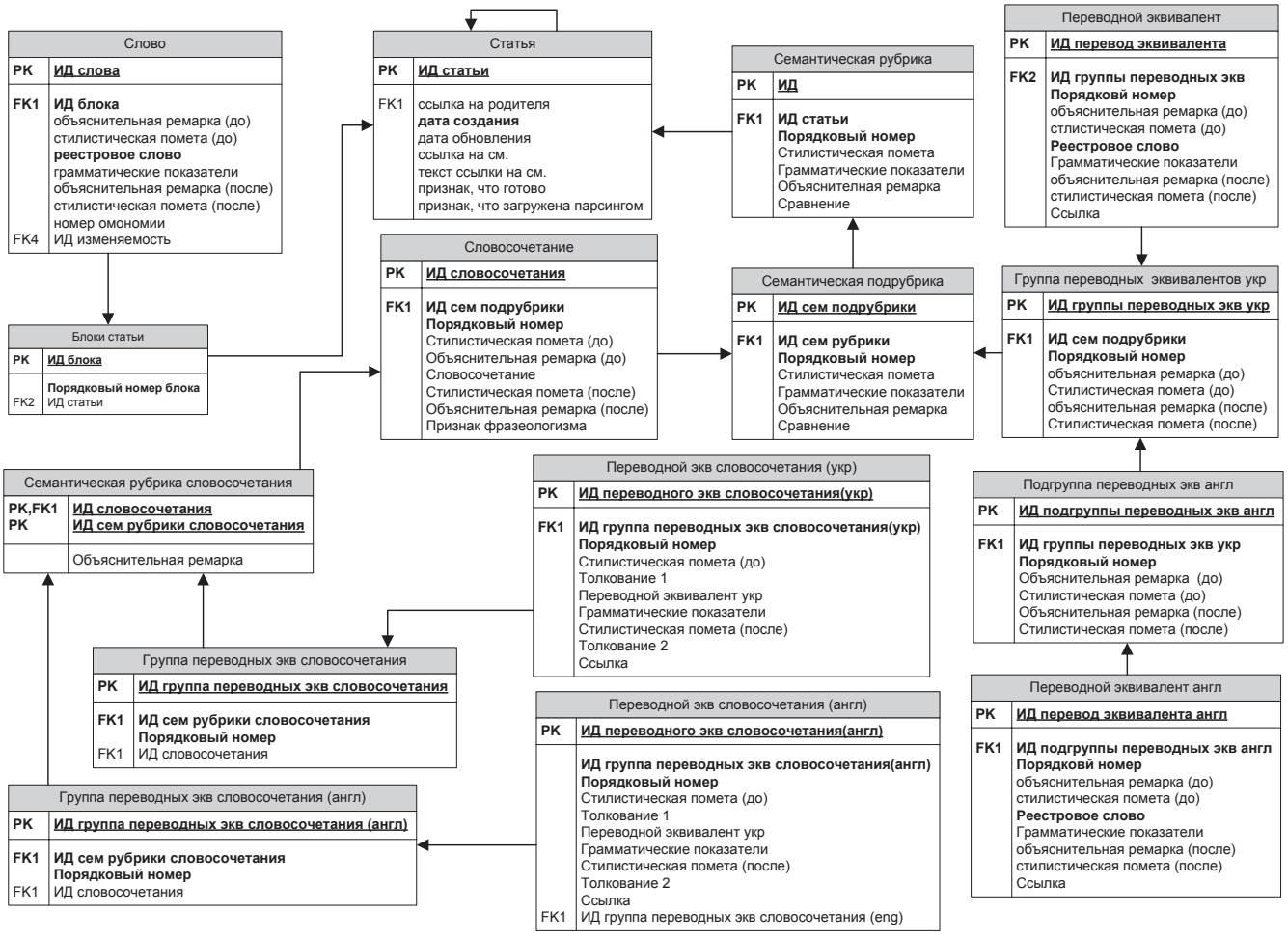

Рис. 1

– для прилагательного – часть речи (*прикм.*), степень сравнения (*вищ. ст.*, *збільш.*, *зменш.*, *зменш.-пестл.*);

– для числительного – часть речи (*числ.*);

– для местоимения – часть речи (*займ.*), разряд (*вказ.*, *пит.*, *особ.*, *означ.*);

– для глагола – вид (*недок.*, *док.*), форма (*акт.*, *безос.*);

– для причастия – часть речи (*дієпр.*);

– для деепричастия – часть речи (*дієприсл.*);

– для наречия – часть речи (*присл.*);

– для предлога – часть речи (*прийм.*);

– для союза – часть речи (*спол.*);

– для частицы – часть речи (*част.*);

– для междометия – часть речи (*виг.*).

На рис. 2, 3 экранными формами представлены поиск в словаре по указанной букве, а также пример редактирования словарной статьи.

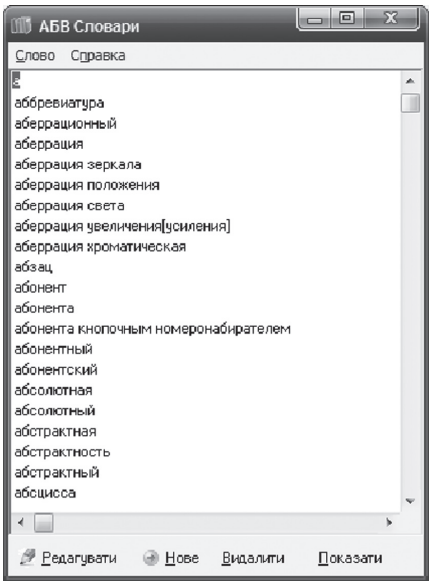

Рис. 2.

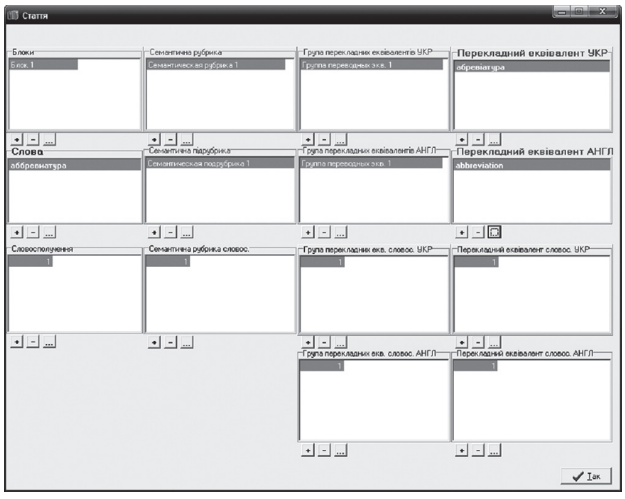

Рис. 3

В перспективе предполагается расширение функциональных возможностей электронного трехязычного словаря (Р → У → А). На данном этапе он позволяет формировать два двуязычных словаря (Р → У, У → А). Для реализации также и остальных (Р → А, У → Р, А → У, А → Р) из всех шести возможных двуязычных словарей ведутся исследования по разработке математической модели с помощью универсального средства описания – алгебры конечных предикатов [3].

3. Схема построения модели трехязычного терминологического словаря на основе трехслойной декомпозиции предикатов

Трехслойная декомпозиция предикатов на примере словоизменения полных непритяжательных имен прилагательных русского языка рассматривалась в [4]. Построенная схемная реализация, обладая свойством обратимости, реализуя широкое распараллеливание обработки информации предиката, получилась обратимой, параллельной и позволяет вычислить результат за несколько тактов. Схема работает в нескольких режимах: вычисляет значение предиката по заданному, определяет неизвестные значения переменных по известным и т.п. Было отмечено, что полученный результат может найти широкое применение в базах данных, где при выполнении некоторых запросов (например, на определение содержания ячеек таблицы) требуется переходить от одной таблицы к другой (достаточно сложный процесс, при котором лавинообразно нарастает информация). Если же от предиката (каждая таблица описывается определенным предикатом) возможно перейти к линейному логическому оператору, то описанное действие выполняется значительно проще и быстрее, а ответ при этом вычисляется (а не находится).

Приведем схематическое описание построения трехслойной декомпозиции предиката, описывающего лексикографическую базу данных терминологического словаря. Для этого необходимо произвести сначала двухслойную декомпозицию соответствующих предикатов 1-го и 2-го рода.

При двухслойной декомпозиции предикатов 1-го рода исходный предикат представляется через свои характеристические функции (сюръекции) и более простой, чем исходный, предикат L (он определен на множестве меньшей мощности).

Рассмотрим тернарный предикат и будем искать для него такое представление, в котором сравнение значений трех соответствующих ему функций f_1, f_2 и f_3 производилось бы с помощью простейшего в некотором смысле предиката. Функциям f_1, f_2 и f_3 соответствуют русский, украинский и английский языки. В нашем случае переменным исходного предиката будут соответствовать все слова и признаки терминологического словаря.

Отметим далее, что множество A_1 отображает значения таблиц лексикографической базы данных, относящихся к русскому языку (множество значений переменной x_1), множество A_2 – значения таблиц, относящихся к украинскому языку, (множество значений переменной x_2), A_3 – значения

таблиц, относящихся к русскому языку (множество значений переменной x_3 соответственно).

Введем понятие сопровождающих эквивалентностей предиката, представив их следующей формулой:

$$E_1(x_1', x_1'') = \forall x_2 \in A_2 \ \ \forall x_3 \in A_3 \ (P(x_1', x_2, x_3) \approx P(x_1'', x_2, x_3)).$$

Пусть далее предикат $P(x_1, x_2, ..., x_n)$ определен на декартовом произведении множеств $A_1 \times A_2 \times ... \times A_n$. Тогда имеет место:

$$E_i(x_i', x_i'') = \forall x_1 \in A_1 \forall x_2 \in A_2 ... \forall x_{i-1} \in$$
$$\in A_{i-1} \forall x_{i+1} \in A_{i+1} ... \forall x_n \in A_n$$
$$(P(x_1, x_2, ..., x_{i-1}, x_i', x_{i+1}, ..., x_n) \approx$$
$$P(x_1, x_2, ..., x_{i-1}, x_i'', x_{i+1}, ..., x_n)).$$

Таким образом,

$$P(x_1, x_2, ..., x_n) = L(f_1(x_1), f_2(x_2), ..., f_n(x_n)),$$

$P(x_1, x_2, x_3) = L(f_1(x_1), f_2(x_2), f_3(x_3))$. Далее находим $L(v_1, v_2, v_3) = P(f_1^{-1}(v_1), f_2^{-1}(v_2), f_3^{-1}(v_3))$, заменив $f_i(x_i) = v_i$.

Построим таблицу предиката $P(x_1, x_2, x_3)$, характеризующую связь между переменной x_1 и отношением (x_2, x_3), т.е. между термином русского языка и переводными эквивалентами на украинском и английском языках. Если совокупность признаков (x_1, x_2, x_3) присутствует в терминологическом словаре, то ставим в таблице единицу, в противном случае – нуль.

Далее аналогичным образом необходимо исследовать связь между признаками русского, английского и терминами украинского языков; а также связи между признаками русского, украинского и терминами английского языков и построить соответствующие классы разбиений.

Двухслойная декомпозиция предикатов 2-го рода дает представление исходного предиката через отображения и простейший предикат, единственный для всех предикатов, подобный предикату равенства. Таким образом, двухслойная декомпозиция предикатов 2-го рода – это следующий шаг к представлению отношения в наиболее общем и простом виде.

Записываем характеристические функции предиката *P*. С этой целью находим характеристические функции $f_i(x_i) = v_i$ эквивалентностей E_i, $(i = \overline{1,3})$.

Записываем образ предиката *P*, описывающего окончание полных непритяжательных имен прилагательных русского языка. С этой целью представим предикат *P* в виде:

$$P(x_1, x_2, x_3) = L(f_1(x_1), f_2(x_2), f_3(x_3)).$$

Откуда находим

$$L(v_1, v_2, v_3) = P(f_1^{-1}(v_1), f_2^{-1}(v_2), f_3^{-1}(v_3)).$$

Составляем таблицы предиката *L*, заменяя признаки именами слоев, в которые они входят, x_i – на v_i, $(i = \overline{1,3})$, *P* – на *L*, затем повторяющиеся столбцы и строки исключаем из таблицы.

Форма записи предиката эквивалентности может быть видоизменена с помощью предиката равенства и характеристических функций. Предикат эквивалентности представим в наиболее общем виде через конъюнкцию своих характеристических предикатов.

Трехслойной декомпозицией предиката E называется его представление в виде:

$$E(x_1, x_2, x_3) = D(g_1^{-1}(f_1(x_1)), g_2^{-1}(f_2(x_2)), g_3^{-1}(f_3(x_3))),$$

где $f_1, f_2, f_3, g_1, g_2, g_3$ – некоторые функции.

Обобщение теоремы об общем виде 2-го рода предиката на *n*-арные предикаты имеет следующий вид:

$$E(x_1, x_2, ..., x_n) =$$
$$= \exists v \in B(F_1(x_1, v) \wedge F_2(x_2, v) \wedge ... \wedge F_n(x_n, v)),$$

где

$$F_i(x_i, v) =$$
$$= \exists x_1 \in A_1 \exists x_2 \in A_2 ... \exists x_{i-1} \in A_{i-1} \exists x_{x+1} \in A_{i+1} ... \exists x_n \in A_n$$
$$S(x_1, x_2, ..., x_n, v),$$

S – функция, присваивающая какие-либо различные имена *v* всем наборам $(x_1, x_2, ..., x_n)$, для которых $E(x_1, x_2, ..., x_n) = 1$; *B* – множество всех таких имен.

Предложенная форма записи общего вида предиката 2-го рода с помощью предиката равенства и отображений более удобна для практики.

Представленный способ нахождения характеристических предикатов с использованием некоторой классифицирующей функции *S*, присваивающей какие-либо различные имена *v* всем парам предметов, для которых предикат равен 1.

Таким образом, получаем

$$E(x_1, x_2, ..., x_n) =$$
$$= \exists v \in B(F_1(x_1, v) \wedge F_2(x_2, v) \wedge ... \wedge F_n(x_n, v)),$$

где $F_1(x_1, v) = \exists x_2 \in A_2 \exists x_3 \in A_3 \ \ S(x_1, x_2, x_3, v)$,

$$F_2(x_2, v) = \exists x_1 \in A_1 \exists x_3 \in A_3 \ \ S(x_1, x_2, x_3, v),$$

$$F_3(x_3, v) = \exists x_1 \in A_1 \exists x_2 \in A_2 \ \ S(x_1, x_2, x_3, v).$$

Предикат $E(x_1, x_2, x_3)$, согласно определению, представляет собой композиции предикатов H_1, H_2, H_3 и D_C:

$$E(x_1, x_2, x_3) = \exists p_1, p_2, p_3 \in C \, (H_1(x_1, p_1) \wedge H_2(x_2, p_2) \wedge$$
$$\wedge H_3(x_3, p_3) \wedge D_C(p_1, p_2, p_3)).$$

4. Анализ и перспективы дальнейших исследований

Уже на первом этапе исследований при проведении декомпозиции предиката 1-го рода становится очевидным, что разработанная лексикографическая база данных не будет достаточной для выполнения свойства «обратимости», т.е. трехязычный словарь не будет работать на всю мощность, предоставляя возможность получать из него не только русско-украинский и украинско-английский, а также другие двуязычные словари и недостающие трехъязычные.

Отметим, что признаки разработанной лексикографической базы данных отображены в 15 таблицах, среди них 7 таблиц («Слово», «Статья», «Семантическая рубрика», «Семантическая под-

рубрика», «Блок статьи», «Словосочетание», «Семантическая рубрика словосочетания») относятся к русскоязычному термину, 4 таблицы («Украинский переводной эквивалент», «Группа украинских переводных эквивалентов», «Украинский переводной эквивалент словосочетания», «Группа украинских переводных эквивалентов») – к переводным эквивалентам украинского и еще 4 («Английский переводной эквивалент», «Группа английских переводных эквивалентов», «Английский переводной эквивалент словосочетания», «Группа английских переводных эквивалентов») – к переводным эквивалентам английского языка.

Непростой является сама структура словарной статьи [1]. На рис. 4 она представлена в графическом виде на примере слова *звучащий* и его переводных эквивалентов.

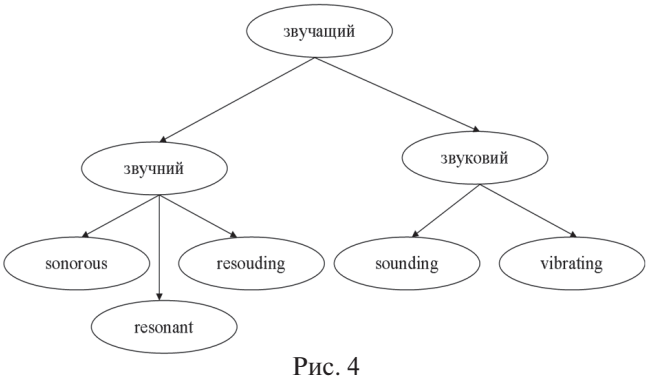

Рис. 4

Приведенная структура представляет собой дерево, главной вершиной которого является слово-термин, на втором уровне находятся узлы, обозначающие семантическую группу, за ней идет уровень семантических подгрупп, только после этого – группа переводных эквивалентов и только потом появляется уровень, который отображает непосредственно сам переводной эквивалент. Таким образом, чтобы получить возможность нахождения всех переводных эквивалентов для реестрового слова, двигаясь снизу вверх, необходимо заполнить все промежуточные поля и признаки. Если в словарь вносить еще и терминологические словосочетания, структура становится еще сложнее.

Выводы

Интеллектуальная обработка информации изначально предполагает детальный анализ предметной области, а также возможностей используемых технических средств. Часто при реализации возникают проблемы, в первую очередь, из-за несоответствия между средствами построения математической модели (неполнота алгебры, как следствие, невозможность расширения/обновления модели – способности к обучению и самообучению) и средствами описания функциональных возможностей самой технической системы. Одним из путей решения этой проблемы является выбор универсального средства описания [5]. Под «универсальностью» математического аппарата подразумевается возможность описания его средствами объекта любой природы.

Таким образом, в статье было описано построенную лексикографическую базу данных терминологического трехязычного словаря. С помощью средств алгебры конечных предикатов был проведен анализ схемы связей между таблицами и структуры самих таблиц, что позволило определить пути решения проблемы создания прямых и обратимых трехязычных электронных словарей.

Использованный метод трехслойной декомпозиции осуществлялся путем обобщения: взяли предикат эквивалентности, убрали свойство однозначности – получили толерантность, убрали еще свойство рефлексивности, заменив его квазирефлексивностью (рефлексивность не на всей области определения, а на ее подмножестве) – получили квазитолерантность. Потом убрали последнее свойство симметричности и получили произвольный n-арный предикат.

Кроме этого, были проанализированы технологические аспекты разработки базы данных терминологического словаря и выбор программных средств для его реализации.

Список литературы: 1. Рабулець, О.Г. Дієслово в лексикографічній системі [Текст] / О.Г. Рабулець, В.А. Широков, К.М. Якименко. – К.: Довіра, 2004. – 259 с. **2.** Остапова, И.В. Лексикографическая структура этимологических словарей и их представление в цифровой среде [Текст] / И.В. Остапова // Прикладная лингвистика и лингвистические технологи: сборник научных трудов. – 2007. – С. 236-245. **3.** Вечирская И.Д. Расслоение предикатов на примере словоизменения прилагательных русского языка [Текст] / И.Д. Вечирская, Г.Г. Четвериков, Т.Н. Федорова // Искусственный интеллект. – 2009. – № 3. – С. 170-177. **4.** Бондаренко, М.Ф. Теория интеллекта [Текст]: учеб. / М.Ф. Бондаренко Ю.П. Шабанов-Кушнаренко. – Харьков: Изд-во СМИТ, 2006. – 571с. **5**. Как переводит компьютер [Электронный ресурс] / С.В.Соколова. – Режим доступа: http://www.translationmemory.ru/technology/ articles/article_Sokolova.php – 05.11.2010 г. – Загл. с экрана.

Поступила в редколлегию 23.05.2011.

УДК 519.7:007.52; 519.711.3

Розробка тримовного термінологічного словника на основі алгебри скінченних предикатів / І.Д. Вечірська // Біоніка інтелекту: наук.-техн. журнал. – 2011. – № 2 (76). – С. 109-113.

Розроблено базу даних термінологічного тримовного словника. Обґрунтовано вибір програмних засобів для його реалізації. Проведено аналіз схеми зв'язків таблиць та їх структури за допомогою методу тришарової декомпозиції предикатів.

Табл. 1. Іл. 4. Бібліогр.: 5 найм.

UDK 519. 519.7:007.52; 519.711.3

About method of linear logical transformations computation / I.D. Vechirskaya // Bionics of Intelligense: Sci. Mag. – 2011. – № 2 (76). –P. 109-113.

The database of terminological three-lingual dictionary is developed. Choosing of software environment to its realization. The connection scheme and structure of tables and is analysed by means of method of predicate three-layer decomposition.

Tab. 1. Fig. 4. Ref.: 5 items.