УДК 519.7

Т. Н. ФЕДОРОВА

О ПОДХОДЕ К ПОСТРОЕНИЮ ЦЕПОЧЕК ЛЕКСИЧЕСКИХ ЕДИНИЦ УКРАИНСКОГО ЯЗЫКА В ЛЕКСИКОГРАФИЧЕСКОЙ СИСТЕМЕ ЭЛЕКТРОННОГО ТОЛКОВОГО СЛОВАРЯ

Рассматривается дальнейшее развитие метода нахождения n-го линейного логического преобразования для построения цепочек в лексикографической системе электронных толковых словарей. Модификация метода характеризуется заданием исходной семантической зависимости на каждом этапе вычисления. Рассматривается реализация метода программой «Побудова гіперланцюгів», которая позволяет строить, редактировать и анализировать цепочки.

1. Введение

В связи с потребностями настоящего времени появляются новые разделы лексикографии, которые подают примеры дальновидных обобщений понятия словаря, а идентификация информационных процессов побуждает к ускорению разработки различных методик формализации языка, а также к созданию все более мощных методов лексикографирования

явлений предметного мира. Пути, по которым движется сегодня лексикография, во многом определяются внешними факторами, среди которых глобализация, становление индустрии знаний и вызванная этим потребность в интеллектуальных средствах экстракции знаний, способных в реальном времени обрабатывать сверхбольшие массивы естественно-речевой информации. В результате сегодня словарное дело переживает особый этап своего развития, находясь под влиянием новых общественных потребностей и новых методов обработки информации, а также используя возможности применения компьютерных технологий при описании и представлении как собственно лингвистической, так и экстралингвистической информации.

Поскольку компьютерная лексикография требует максимальной формализации своего объекта, возникла проблема в углублении содержания и формализации самого понятия словаря как специфического культурно-информационного объекта в процессе развертывания фундаментальных для языка отношений «субъект - объект » и « форма-содержание ». Ответом на эту потребность стала разработка лексикографических систем, которые поясняют внутренние механизмы, побуждающие к приобретению естественно-речевой информацией словарной формы.

Одна из проблем лексикографии касается применения словарей в формировании лингвистических компонент концептографических систем представления знаний и использовании их в средствах экстракции знаний. Это требует не только больших словарных массивов, но и побуждает к нахождению в традиционных словарных текстах скрытых семантических структур. С другой стороны, онто- и концептографическая проблематика требует более активной и содержательной интеграции лексикографии в современную индустрию знаний [1].

Целью данной работы является модификация метода линейного логического преобразования n-й степени для построения цепочек лексических единиц украинского языка[2, 3]. Это позволит повысить скорость и точность обработки словарных статей с помощью анализа отношений толкования и построения гиперцепочек между лексическими единицами украинского языка. Для достижения поставленной цели необходимо решить следующие задачи: усовершенствовать метод, построить алгоритм и программно реализовать его.

2. Метод построения цепочек лексических единиц украинского языка

Цепочки вида «толкуется через» используют для построения систем, которые могли бы находить в тексте или в его фрагменте не только конкретно заданное слово, но и это слово по его содержанию, описанию. Примерами таких систем являются «ПроСеКа» [4] и «СКАЗКА-2» [5].

В работе [2] изложен и обоснован метод нахождения n-го линейного логического преобразования. Было доказано утверждение о том, что при нахождении степени линейного логического преобразования, если на двух последующих шагах значение преобразования повторяется, то такое линейное преобразование и будет искомым. Этот же критерий нахождения n-го линейного логического преобразования был использован при решении задачи построения цепей лексических единиц.

Рассмотрим в общем виде метод нахождения n-го линейного логического преобразования. $P(x)$, $Q(y)$ – предикаты, $K(x,y)$ – ядро линейного логического преобразования, M – множество, элементы которого являются логическими векторами. Линейные логические преобразования можно представить в виде $Q(y) = \exists x \in M(K(x,y)P(x))$.

Приведем в общем виде формулу вычисления преобразования $P^{(n)}(x)$ и $Q^{(n)}(y)$ в зависимости от предикатов $P(x)$ и $Q(y)$ соответственно.

Пусть $Q(y) = K(x,y)P(x)$, $P'(x) = K(y,x)Q(y)$. Преобразование из $P'(x)$ представим в следующем виде:

$$Q'(y) = K(x,y)P'(x) \overset{(1)}{=} K(x,y)K(y,x)Q(y) = KQ(y),$$

$$P'(x) \overset{(1)}{=} K(y,x)K(x,y)P(x) = K'P(x).$$

Преобразование из $P''(x)$ представим в виде:

$$Q''(y) = K(x,y)P''(x) = K(x,y)K(y,x)Q'(y) \overset{(3)}{=} KKQ(y),$$

$$P''(x) = K(y,x)Q'(y) = K(x,y)K(x,y)P'(x) \overset{(4)}{=} K'K'P(x).$$

Действуя аналогичным способом, получаем формулу вычисления n-го линейного логического преобразования вида:

$$Q^{(n)}(y) = \bigwedge_{i=1}^{n} K_i Q(y), \text{ где } K_i = K = K(x,y)K(y,x),$$

$$P^{(n)}(x) = \bigwedge_{i=1}^{n} K_i' P(x), \text{ где } K_i' = K' = K(y,x)K(x,y).$$

Если линейные логические преобразования n-й и $n+1$-й степени совпадают, то n-е логическое преобразование далее не изменится, оно стабилизируется на n-м шаге.

Алгебра конечных предикатов позволяет формализовать подход к построению цепочек следующим образом: пусть $P(x)$ – слово, $K(x_{n-1},x_n)$ – семантическая зависимость, определяющая функцию толкования, M – множество всех слов в словарных статьях электронного толкового словаря.

Задаем $P(x_1)$, $K_1(x_1,x_2)$. Вычисляем $P(x_2)$:

$$P(x_2) = \exists x_1 \; P(x_1) K_1(x_1, x_2). \quad (1)$$

Задаем $K_2(x_2,x_3)$. Вычисляем $P(x_3)$:

$$P(x_3) = \exists x_2 \; P(x_3) K_2(x_2, x_3). \quad (2)$$

Задаем $K_3(x_3,x_4)$. Вычисляем $P(x_4)$:

$$P(x_4) = \exists x_3 \; P(x_3) K_3(x_3, x_4) \quad (3)$$

и т.д. Вычисление будет проходить до того момента, пока не выполнится условие завершения построения цепочки:

$$x_n = x_i \text{ если } \exists \; K_i(x_i, x_n), \; i = \overline{1, n-1}. \quad (4)$$

Приведем в общем виде формулу вычисления преобразования $P(x_n)$:

$$P(x_n) = \exists x_{n-1} \; P(x_{n-1}) K_{n-1}(x_{n-1}, x_n). \quad (5)$$

На рис. 1 изображено графическое представление метода построения цепей лексических единиц, где x – слово, $x(i,j)$, i – номер уровня, j – индекс слова на уровне i.

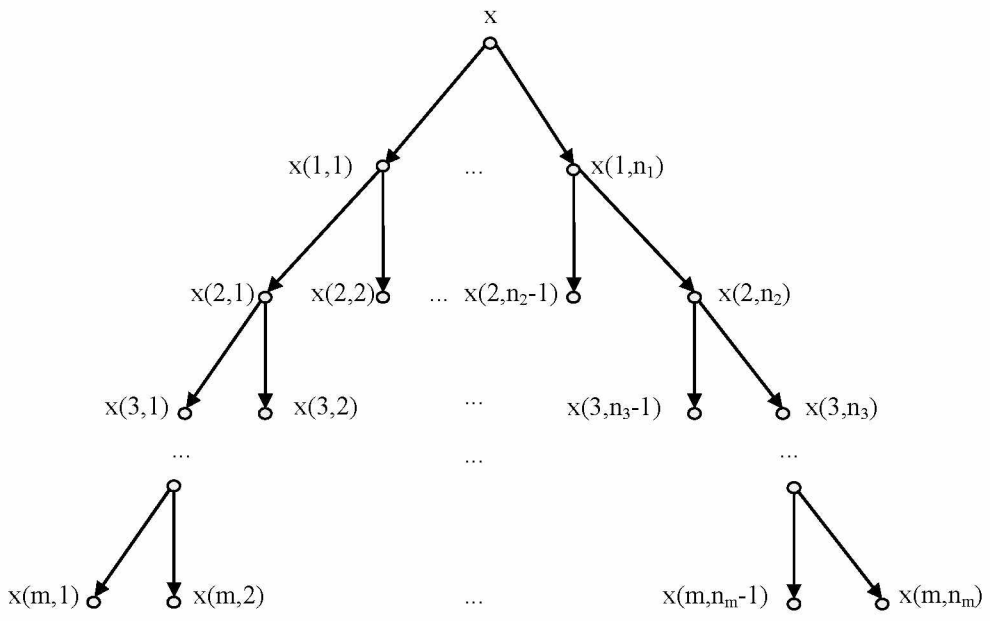


Рис. 1. Графическое представление метода

Рассмотрим пример вычисления линейного логического преобразования метода построения цепочек лексических единиц для слова «Абажур». Цепочка «Абажур -> прилад -> пристрій -> обладнання -> прилад».

«Абажур – **частина світильника**, звичайно у вигляді ковпака, **прилад** призначений для зосередження і відбиття світла та захисту очей від його впливу.»

«Прилад – 1. **Інструмент**, **предмет**, який використовується для виконання певної дії. 2. Спеціальний **пристрій**, призначений для певної мети (вимірювання чого-небудь, управління чимось, контролю, спостереження за чим-небудь і т. ін.). 3. Сукупність відповідних інструментів, предметів, необхідних для виконання певної роботи.»

«Пристрій – **пристосування**, **обладнання**, за допомогою якого виконується яка-небудь робота або спрощується, полегшується певний виробничий процес.»

«Обладнання – сукупність **механізмів**, **приладів**, необхідних для чого-небудь; **спорядження**.»

Задаем

$$P(x_1) = \{x_1^{\text{абажур}}\},$$

$$K_1((x_1^{\text{абажур}} x_2^{\text{частина}}) \vee (x_1^{\text{абажур}} x_2^{\text{світильник}}) \vee (x_1^{\text{абажур}} x_2^{\text{прилад}})),$$

вычисляем

$$P\left(x_2^{\text{прилад}}\right) = x_1^{\text{абажур}} \wedge x_1^{\text{абажур}} x_2^{\text{прилад}}.$$

Задаем

$$K_2((x_2^{\text{прилад}} x_3^{\text{інструмент}}) \vee (x_2^{\text{прилад}} x_3^{\text{предмет}}) \vee (x_2^{\text{прилад}} x_3^{\text{пристрій}})),$$

вычисляем

$$P\left(x_3^{\text{пристрій}}\right) = x_2^{\text{прилад}} \wedge x_2^{\text{прилад}} x_3^{\text{пристрій}}.$$

Задаем

$$K_3((x_3^{\text{пристрій}} x_4^{\text{пристосування}}) \vee (x_3^{\text{пристрій}} x_4^{\text{обладнання}}))$$

вычисляем

$$P\left(x_4^{\text{обладнання}}\right) = x_3^{\text{пристрій}} \wedge x_3^{\text{пристрій}} x_4^{\text{обладнання}}$$

Задаем

$$K_4((x_4^{\text{обладнання}} x_5^{\text{механізм}}) \vee (x_4^{\text{обладнання}} x_5^{\text{прилад}}) \vee (x_4^{\text{обладнання}} x_5^{\text{спорядження}}))$$

вычисляем

$$P\left(x_5^{\text{прилад}}\right) = x_4^{\text{обладнання}} \wedge x_4^{\text{обладнання}} x_5^{\text{прилад}}$$

На 5-м шаге выполнится условие завершения построения цепочки:

$$x_5^{\text{прилад}} = x_2^{\text{прилад}} \text{ для } \exists K_2(x_2^{\text{прилад}} x_5^{\text{прилад}}).$$

3. Программная реализация метода построения цепочек лексических единиц

На основе метода построения цепочек лексических единиц для украинского языка разработана программа «Побудова гиперланцюгів», которая предназначена для описания семантических отношений между лексическими единицами естественного языка.

Для построения гиперцепей, которые связываются отношением «толкуется через», используется толковый словарь «Виртуальной лексикографической лаборатории Украинского языково-информационного фонда» [3]. Целью системы является автоматизация обработки текстов с помощью анализа отношений и построения цепочек между лексическими единицами украинского языка. С помощью программы возможно строить, редактировать и анализировать цепочки.

Программа позволяет выбирать путь проведения поиска по базе гиперцепочек или по электронному толковому словарю украинского языка. Программа включает две базы, в одной хранятся цепочки, разработанные вручную, а в другой - цепочки, которые строятся на базе электронного толкового словаря. Если для слова уже была построена цепочка с помощью электронного словаря, программа будет выводить результат, ранее сохраненный

в базе, что позволяет сократить время их построения. Если же пользователю необходимо построить гиперцепочку заново, то задается параметр «принудительный поиск», и программа будет строить заново цепочку, обращаясь только к электронному словарю. Есть возможность одновременного поиска по двум базам слов. Результат выводится в отдельных окнах так, что пользователь может сравнивать полученные результаты (рис. 2).

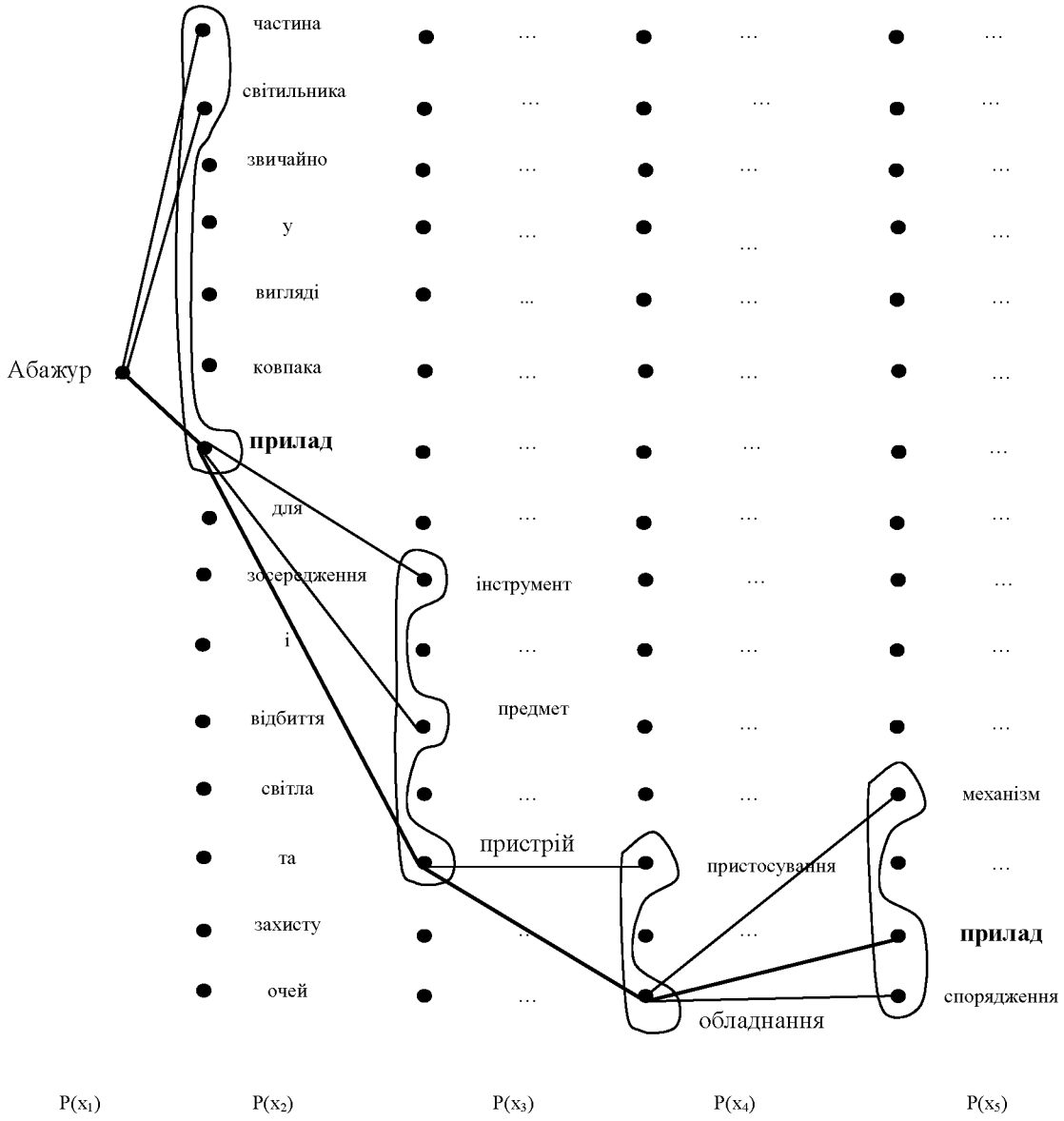


Рис. 2. Графическое представление одного из вариантов построения цепочек для слова «Абажур»

При построении гиперцепочки слова и по словарю, и по базе выводится комбинированный результат (рис. 3).

Программа позволяет контролировать каждый шаг ее выполнения, для этого необходимо установить отметку возле поля «Подтверждать каждый шаг».

Каждое слово в гиперцепочке можно редактировать следующим образом: добавить или удалить слово, для каждого слова отдельно можно отобразить дочерние слова (рис. 4). Для вызова окон редактирования слов нужно кликнуть на нужном слове и в контекстном меню выбрать вид редактирования (рис. 5).

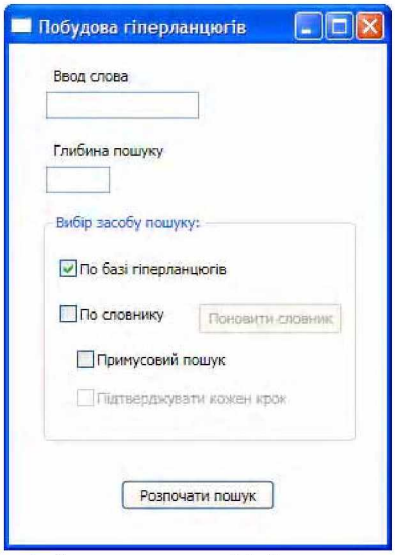
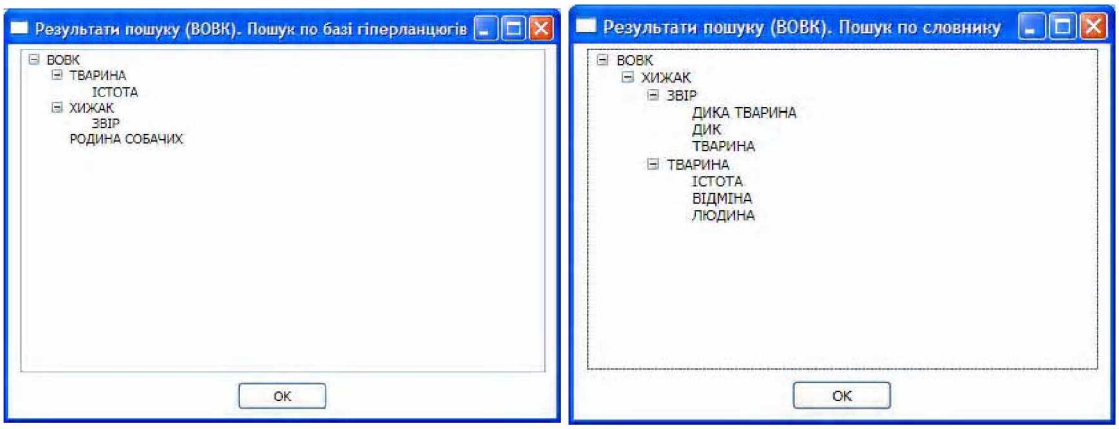


Рис. 3. Интерфейс программы «Побудова гіперланцюгів»



а б

Рис. 4. Результаты построения гиперцепочки для слова «вовк»:

а – по базе гиперцепочек; б – по словарю

Алгоритм, выполняемый для автоматизации поиска слов в электронном словаре:

1. Поиск словарной статьи. Слово, для которого выполняется поиск, вводится в поле «Найти слово» электронного словаря. Вызывается словарная статья путем вызова действия кнопки «Найти».

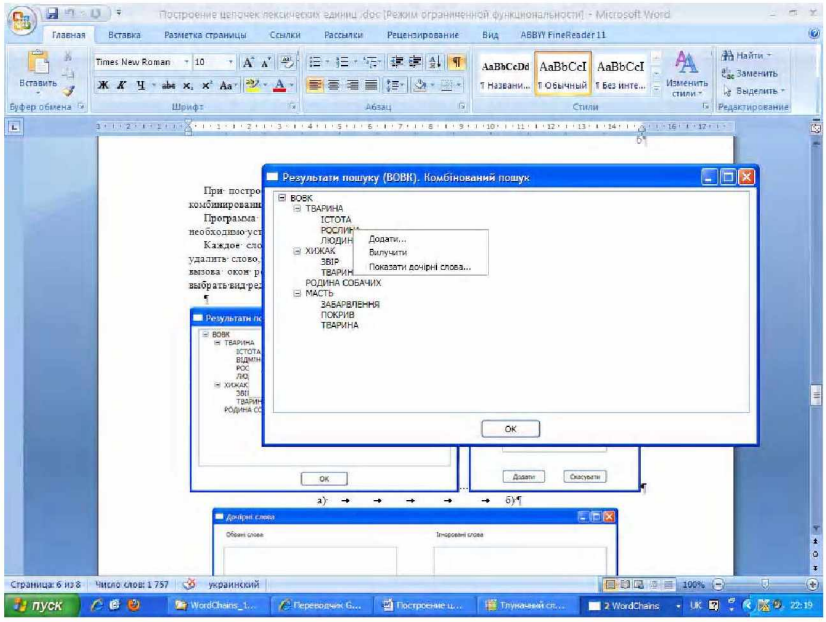
2. Анализ результата поиска. Из полученной статьи выделяются слова, дополнительные свойства слова, является ли слово существительным, и выделяется сама словарная статья, которая разделяется на отдельные слова.

3. Сбор информации о каждом слове словарной статьи. Проверяется, есть ли найденное слово в локальной базе данных. В случае, если слово отсутствует, вызывается новая соответствующая статья.

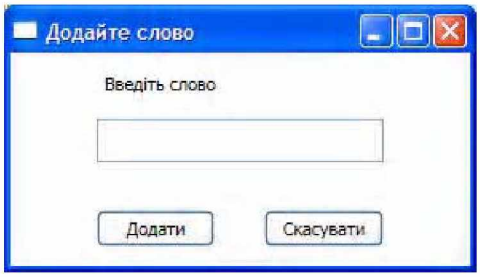
4. В новой статье выбираются слова. В случае если выбранное слово является существительным – оно добавляется в локальную базу данных, если нет – в список игнорируемых слов.

5. Вызывается кнопка «Назад» для возврата к основной статье.

6. Повторяется пункт 3 для каждого слова статьи.



а



б

в

Рис. 5. Меню для вызова окна редактирования слова (а); окно добавления слов (б); окно редактирования дочерних слов (в)

4. Выводы

Научная новизна: получил дальнейшее развитие метод нахождения n-го линейного логического преобразования для построения цепочек в лексикографической системе электронных толковых словарей путем задания исходной семантической зависимости на каждом этапе вычисления. Также рассмотрена реализация метода программой «Побудова гіперланцюгів», которая позволяет строить, редактировать и анализировать цепочки.

Практическое значение: метод позволяет распараллелить процесс обработки словарных статей с помощью анализа отношений толкования и построения гиперцепей между лексическими единицами украинского языка.

Перспективы исследования: нахождение в традиционных словарных текстах скрытых семантических структур, что позволит создавать более мощные методы лексикографирования явлений предметного мира.

Список литературы: 1. *Широков В. А.* Комп'ютерна лексикографія. К.: Наук. думка, 2011. 351 с. **2.** *Вечирская И.Д.* Линейные логические преобразования и их применение в исскуственном интеллекте: Автореферат дисс. канд. техн. наук. Х., 2007. 28с. **3.** *Широков В.А.* Лінгвістичні та технологічні основи тлумачної лексикографії. К.: Довіра, 2010. 295 с. **4.** *Рафаева А. В.* Программа семантической классификации лексики «ПроСеКа» // Материалы международной научной конференции "Горизонты прикладной лингвистики и лингвистических технологий" (MegaLing'2009). 20-27 сентября 2009, Украина, Киев.С. 67. **5.** *Рафаева А.В.* Использование программы ПРОСЕКА в исследовании сказок // Прикладна лінгвістика та лінгвістичні технології : MegaLing-2009 : Зб. наук. пр. / НАН України. Укр. мовн.-інформ. фонд, Таврійський нац. ун-т ім. В.І.Вернадського; За ред. В.А.Широкова. К. : Довіра, 2010. С. 378–382.

Поступила в редколлегию 12.06.2012

Фёдорова Татьяна Николаевна, аспирантка кафедры програмной инженерии ХНУРЭ. Научные интересы: математическая и прикладная лингвистика, алгебра логики. Интересы: изучение иностранных языков, спортивные бальные танцы, балет, вышивка, конный спорт, катание на коньках. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. 702-14-77, E-mail: tanja_fyodorova@mail.ru.