КОМП'ЮТЕРНІ ЗАСОБИ, МЕРЕЖІ ТА СИСТЕМИ

УДК 004. 415

А.В. ПАЛАГИН, Н.Г. ПЕТРЕНКО, М.П. СЛАБКОВСКАЯ

ОБ ОДНОМ ПОДХОДЕ К РАЗРАБОТКЕ АППАРАТНЫХ СРЕДСТВ ПОДДЕРЖКИ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА

A.V. Palagin, N.G. Petrenko, M.P. Slabkovska

ABOUT ONE APPROACH TO THE DEVELOPMENT OF THE HARDWARE SUPPORT FOR LINGUISTIC ANALYSIS

Designed hardware morphological processor, which provides higher performance for procedures corpus linguistic analysis of large amounts on the order of 2 or more.

Key words: linguistic analysis, hardware morphological processor.

Розроблено апаратний морфологічний процесор, який забезпечує підвищення продуктивності при виконанні процедур лінгвістичного аналізу корпусів текстів великого обсягу на 2 і більше порядки.

Ключові слова: лінгвістичний аналіз, апаратний морфологічний процесор.

Разработан аппаратный морфологический процессор, обеспечивающий повышение производительности при выполнении процедур лингвистического анализа корпусов текстов большого объема на 2 и более порядка.

Ключевые слова: лингвистический анализ, аппаратный морфологический процессор.

Введение. Обработка речевой или текстовой информации обеспечивается лингвистическим процессором, будь-то на уровне "языкового" сознания человека или в компьютерной системе (КС). В КС он является основной компонентой, реализующей распознавание и понимание входной естественно-языковой информации, извлечение из нее первичных знаний с их последующим формально-логическим представлением. Полученной информационной структуры уже достаточно для реализации (знание-ориентированных) процедур для решения прикладных задач, принятия решений и т. п.

Постановка задачи. Одной из важных задач на пути разработки общей теории компьютерной обработки предметных знаний, представленных в естественно-языковой форме, является построение эффективных лингвистических процессоров. Эта задача особенно актуальна для приложений обработки лингвистических корпусов текстов (ЛКТ) сверхбольших объемов (и в реальном времени). Благодаря тому, что современные персональные компьютеры программным способом выполняют лингвистический анализ одного слова входного текста средней длины примерно за одну миллисекунду, а такой анализ занимает значительную часть компьютерного времени в общей лингвистической обработке. При этом время обработки входного текста даже относительно небольшого объема занимает от нескольких до десятков минут. В итоге для приложений, работающих в реальном времени, часть информации

будет потеряна.

Поэтому задача существенного (на 2 порядка и более) повышения быстродействия лингвистического анализа является актуальной.

Следует отметить, что указанное повышение быстродействия может быть достигнуто за счет дополнительных аппаратурных затрат как стандартной, так и специальной разработки. Аппаратные средства (АС) первого типа являются продуктом известных фирм, доступны на рынке и к ним прилагается САПР. Несомненным лидером таких АС на рынке являются платы с установленными на них ПЛИС (в которых есть сверхбыстродействующая память) и быстродействующая память большого объема [2]. АС второго типа являются специализированной разработкой, для них необходимо спроектировать архитектурно-структурную организацию процессора, электрическую схему или граф-схемы алгоритмов, специальное программное обеспечение управления ими и драйверы совмещения с операционной системой компьютера. С точки зрения реализации лингвистического процессора оба эти варианта АС имеют свои преимущества и недостатки. Для АС первого типа преимуществом является то, что они доступны на рынке, их вычислительная мощность постоянно увеличивается разработчиками, к ним уже прилагается программное обеспечение, а проект аппаратного лингвистического процессора (АЛП) может быть разработан за время от 2-х месяцев. Недостатком этих АС является малый процент использования установленного на плате оборудования. К преимуществам АС второго типа следует отнести повышение быстродействия на 1-2 порядка по сравнению с АС первого типа, что является главным критерием при разработке АЛП. А к недостаткам – необходим коллектив разработчиков (системотехников и программистов), и время разработки проекта оценивается от 1 года.

Повышение быстродействия реализации алгоритма лингвистического анализа для обоих типов АС достигается за счет перевода операторов алгоритмического и программного уровня (реализация лингвистического анализа программным способом) на нижние уровни интерпретации (согласно логико-информационной модели [1]): для АС первого типа – на микропрограммный уровень, для АС второго типа – на микропрограммный и частично на физический уровни.

В работе [1] приведены дополнительные доводы целесообразности реализации ЛП в целом, и морфологического процессора в частности, аппаратными средствами. Например, аппаратная реализация предоставляет возможность параллельной обработки всех слов одного предложения одновременно. При этом упрощаются алгоритмы синтаксического и семантического анализа.

Следует отметить важную особенность реализации АЛП с применением ПЛИС-технологии, для которой существует возможность реконфигурации структуры АЛП на физическом уровне как инструменте настройки (перестройки) на обработку ЛКТ заданной ПдО или решения задачи предметно-проблемной ориентации аппаратных средств. При этом на логических схемах ПЛИС построены так называемые адаптивные логические сети (АЛС) с памятью [2].

Основная часть. Лингвистический процессор (аппаратный или программный) интерпретирует текстовую информацию (некоторый естественно-языковый объект (ЕЯО) – документ, статья, монография или ЛКТ) в соответствии с этапами лингвистического анализа: графематического, морфологического, синтаксического и семантического (точнее, поверхностно-семантического). Результат работы АЛП – информационная структура, предназначенная для проведения глубинно-семантического анализа в подсистеме экстралингвистической обработки, задачей которой является извлечение и формирование структуры понятий, т. е. автоматическое извлечение из ЕЯО знаний, реализация их прагматической интерпретации в терминах прикладной задачи или соответствующая реакция, присущая человеку.

На рис. 1 показана функциональная схема такого АЛП, включающая соответствующие подсистемы реализации этапов лингвистического анализа и языково-онтологическую картину мира (ЯОКМ), использование которой является одним из основных отличий АЛП от классического лингвистического процессора, а ее проектирование рассмотрено в [1].

На рис. 1 приняты следующие сокращения:

АСС – общепринятые аббревиатуры, сокращения и специальные символы;

ГфА – графематический анализ;

МА – морфологический анализ;

СнА – синтаксический анализ;

СА – семантический анализ;

Ψ – процедура отображения граф-схемы алгоритма реализации соответствующей подсистемы лингвистического анализа в соответствующую сеть комбинационных автоматов с памятью (*TnM*) в терминах САПР ПЛИС;

ЛЕ – лексема;

МХ – морфологическая характеристика;

О – объект; Д – действие; ХО – характеристика объекта; ХД – характеристика действия.

Исходной информацией для АЛП является ЛКТ заданной ПдО. Он включает конечное множество текстов $\{T_k\}, k = \overline{1, K}$, K – количество текстов в ЛКТ, которые последовательно поступают на вход подсистемы графематического анализа.

В процессе выполнения общего алгоритма лингвистического анализа текст $\{T_k\}$ поэтапно преобразуется в графематическую, морфологическую, синтаксическую и семантическую структуры, каждая из которых имеет свою модель представления. В работе рассмотрена только подсистема морфологического анализа.

На рис. 2 показана блок-схема подсистемы морфологического анализа, а на рис. 3 – диаграмма состояний с описаниями исполняемых процедур и анализируемыми условиями.

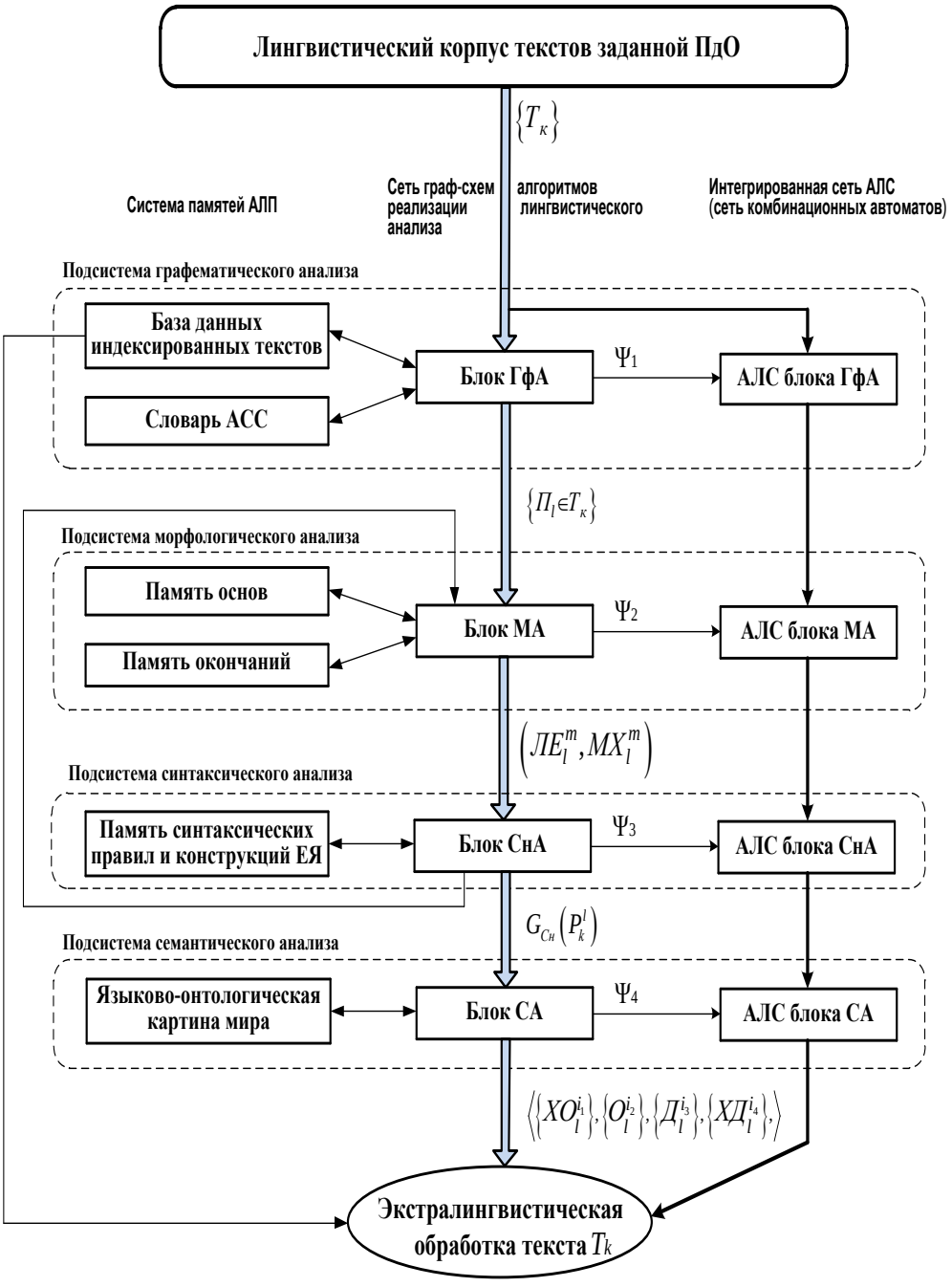

РИС. 1. Функциональная схема АЛП

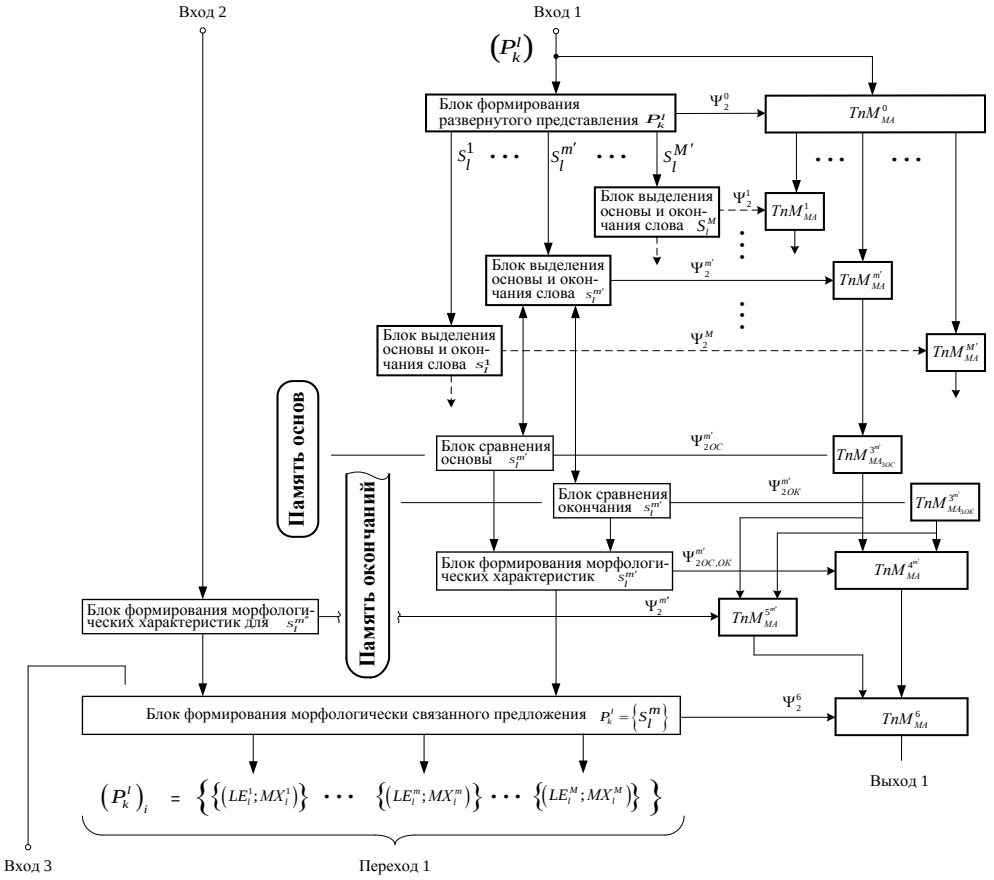

РИС. 2. Функциональная схема подсистемы МА

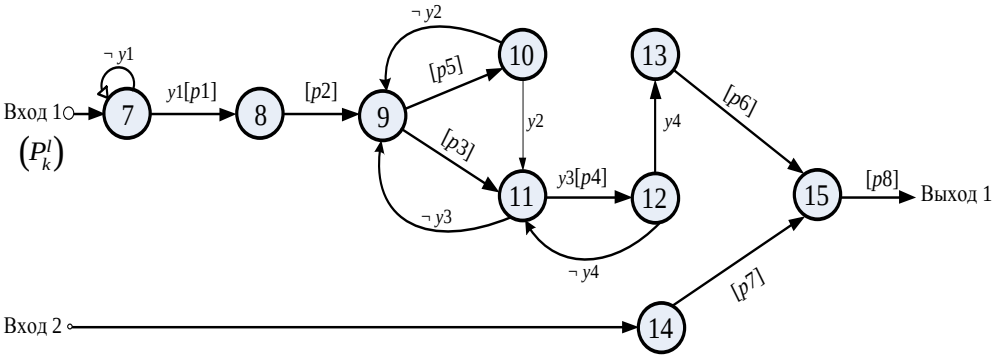

РИС. 3. Диаграмма состояний подсистемы МА

$p1$ – первая процедура для словоформ P_k^l, не являющихся АСС. Формирует развернутое представление предложения P_k^l;

$p2$ – формирует раздельное представление основы и окончания словоформы $S_l^{m'}$. Выполняет предварительные установки информационных и управляющих регистров;

$p3$ – выполняет сравнение основы словоформы $S_l^{m'}$ с содержимым памяти основ;

$p4$ – формирует список основ-омонимов;

$p5$ – выполняет сравнение окончания словоформы $S_l^{m'}$ с содержимым памяти окончаний;

$p6$ – формирует лексемы и морфологические характеристики $S_l^{m'}$, в том числе и омонимов;

$p7$ – формирует морфологические характеристики $S_l^{m''}$;

$p8$ – формирует морфологически связанные предложения P_k^l, раздельные для всех неоднозначных словоформ;

$y1$ – условие начала приема последовательности словоформ предложения P_k^l;

$y2$ – условие сравнения окончания словоформы $S_l^{m'}$;

$y3$ – условие сравнения основы словоформы $S_l^{m'}$;

$y4$ – условие составления полного списка основ-омонимов.

В аппаратном морфологическом процессоре (АМП) для обработки каждой словоформы $S_l^{m'}$ предложения P_k^l выделен отдельный аппаратный блок, в котором из словоформы выделяются основа и окончание, причем принципы такого выделения отличаются от традиционных и описаны в [3]. Совокупность таких блоков, параллельно обрабатывающих все словоформы предложения P_k^l, составляют одну из основных компонент подсистемы морфологического анализа. Максимальное количество блоков определяется на основе статистических характеристик заданного ЛКТ, в частности, параметра максимального количества вхождений словоформ в предложения.

Выделенные основа и окончание словоформы $S_l^{m'}$ поступают соответственно в блок сравнения основы и блок сравнения окончания, в которых по ассоциативному принципу формируется адрес фрагмента ячеек памяти основ и окончаний, в котором хранится морфологическая структура словоформы $S_l^{m'}$. Подробный алгоритм морфологической обработки одной словоформы и его аппаратная реализация описаны в [4].

Аналогичным образом формируются морфологические характеристики для всех словоформ $\left\{S_l^{m'}\right\}$ предложения $\boldsymbol{P}_k^l$. Следует отметить, что указанные морфологические характеристики учитывают особенности только текстов научно-технического и делового стилей.

В подсистеме МА выполняется параллельный анализ словоформ $\left\{S_l^{m'}\right\}$, принадлежащих предложению $P_k^l \in T_k$, где $m' = \overline{1, M}, M$ – количество словоформ в предложении $\boldsymbol{P}_k^l$, $l = \overline{1, L}$, L – количество предложений в тексте $\boldsymbol{T}_k$. На выходе подсистемы в блоках формирования морфологических характеристик формируется выходная информация $\left\{Л_l^{m'}, MX_l^{m'}\right\}$, к которой присоединяется морфологическая информация о присутствующих в анализируемом предложении АСС $S_l^{m''}, m'' \notin M$. Причем, последнее множество может быть и пустым. На выходе подсистемы МА формируется морфологическая структура вида $\left\{\left(ЛЕ_l^1, MX_l^1\right),\ldots, \left(ЛЕ_l^m, MX_l^m\right),\ldots, \left(ЛЕ_l^M, MX_l^M\right)\right\}$, которая является входной информацией для подсистемы синтаксического анализа.

В заключение работы АЛП будет сформирована таблица шаблонов предложений текста $\left\{T_k\right\}$, в которой хранится информационная структура для всего текста в целом, и которая является входной информацией для подсистемы экстралингвистической обработки текста $\left\{T_k\right\}$. В ней выполняется формально-логическое представление (перевод) предложений и текста в целом в подходящей формальной теории первого порядка, например, сначала в модифицированные концептуальные графы [1] и затем в логику предикатов первого порядка.

Структурная организация и проектирование АМП

Далее описана аппаратная реализация подсистемы морфологического анализа (или АМП), причем только последовательного анализа словоформ входного предложения. Для реализации параллельной обработки всех словоформ предложения потребуется *K* блоков морфологического анализа, где *K* – максимальное количество вхождений словоформ в предложение.

Общая схема реализации МА (независимо от способа реализации) сводится к приему последовательности слов, составляющих входной текст, распознаванию или дешифрации анализируемого слова и нахождения соответствующей ему так называемой “точки в гиперпространстве” (или реализация табличного метода анализа), в которой анализируемому слову приписаны все необходимые морфологические характеристики. Это пространство представляет собой по осям X_i части речи заданного ЕЯ, где $i = \overline{1, n}$, n – количество частей речи, а по осям Y_i – последовательность словоформ *i*-ой части речи.

Описанная выше последовательность шагов МА является “идеальной” и

практически нереализуемой для современных микроэлектронных технологий, а приближение к ней возможно только для аппаратной реализации алгоритма МА. Для “идеальной” реализации понадобился бы дешифратор (или память) с адресацией 2^{256} разрядов. Этот параметр определен из того, что для кодирования одной буквы (символа) слова требуется 8 бит, а максимальное количество символов самых длинных словоформ в ЛБД общеупотребительной лексики украинского языка “Словники України” равно 32. Отсюда и получена степень двойки (8х32 = 256).

Классический программный МА выполняется последовательно по буквам, начиная с окончания, нахождения основы словоформы и формирования последовательности омонимов анализируемого слова. При этом для каждого омонима формируется свое множество морфологических характеристик.

Отметим, что в общем случае только “идеальная” аппаратная реализация позволяет избежать раздельного анализа окончания и основы словоформы.

Таким образом, если условно расположить на плоскости по оси *X* реализацию МА классическим программным способом, то описанная выше “идеальная” аппаратная реализация будет расположена по оси *У*, а все другие реализации ($T_г$, $Q_г$) будут расположены между ними.

Обобщенная схема АС реализации алгоритма МА для некоторого решения $T_г$, $Q_г$ показана на рис. 4. На нем приняты следующие обозначения:

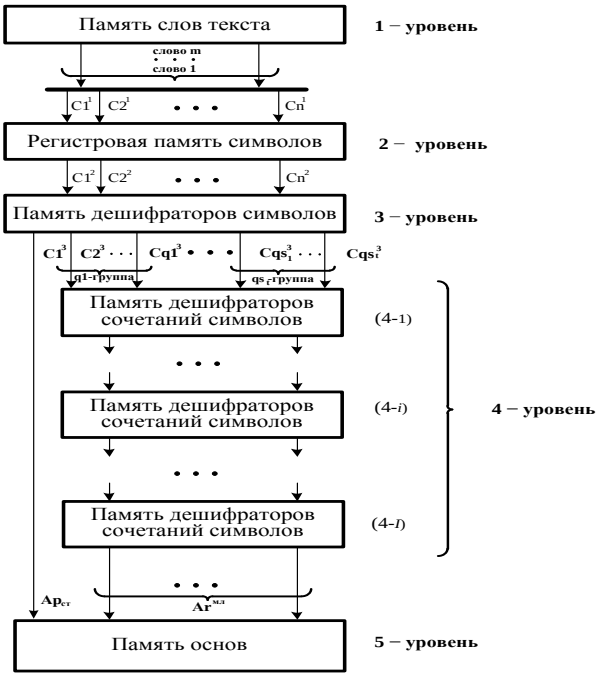

РИС. 4. Обобщенная схема аппаратных средств МА

m – количество слов в анализируемом тексте. Эта последовательность формируется на этапе графематического анализа, записывается в память слов текста и является исходными данными для МА;

$C_1^1, C_2^1, ..., C_n^1$ – максимальное количество букв (символов) в словах анализируемых текстов;

$C_1^3, C_2^3, .., C_{q1}^3, ..., C_{qs_1}^3, ..., C_{qs_t}^3$ – первая буква слова и qs групп сочетаний символов (начиная со второго), которые формируются на основе статистических характеристик и заданных ограничений на оборудование;

A_p^{cm} – старшие разряды адреса памяти слов;

$A_r^{мл}$ – младшие разряды адреса памяти слов.

Суть подхода к построению схемы заключается в "усечении" адресного пространства, необходимого для "идеальной" реализации, до адресного пространства памяти, представленной на стандартном оборудовании. Для этого служит 4-ый уровень на рис. 4.

Выводы. Моделирование описанного АМП выполнено в САПР ПЛИС Xilinx ISE 8.2i с использованием платы HTG-V4PCIE, на которой установлены следующие аппаратные средства, доступные для пользователя и необходимые, в частности, для практической реализации АМП: 1) кристалл ПЛИС Virtex-4, который содержит 376 блоков СОЗУ 18Kbx1, с возможностью организации от 16Kbx1 до 512x36 бит (www.hilinx.com/pro-ducts/boards_kits/virtex6.htm); 2) внешняя (по отношению к кристаллу ПЛИС) память RAM – два независимых блока 64Mx16 бит на которых реализованы памяти основ и окончаний.

Моделирование показало, что разработанный АМП обеспечивает повышение на 2 и более порядка производительности при выполнении процедур лингвистического анализа корпусов текстов большого объема, что особенно важно в системах реального времени. Полученный эффект основан на учете статистических характеристик полного множества словоформ данного языка (при построении архитектуры и структуры лингвистического процессора).

1. *Палагин А.В., Крывый С.Л., Петренко Н.Г.* Онтологические методы и средства обработки предметных знаний. – [Монография]. – Луганск: изд-во ВНУ им. В. Даля, 2012. – 324 с.
2. *Палагин А.В., Опанасенко В.Н.* Реконфигурируемые вычислительные системы. – К.: Просвіта, 2006. – 280 с.
3. *Петренко М.Г., Палагін О.В., Величко В.Ю., Кривий С.Л.* Розробка методів та засобів онтолого-лінгвістичного аналізу природномовних об'єктів. – Київ: Ін-т кібернетики НАН України, 2009. – 38 с. – (Препринт / НАН України, Ін-т кібернетики ім. В. М. Глушкова; 2009-2).
4. *Пристрій* для морфологічного аналізу природномовних текстів / [Палагін О.В., Петренко М.Г., Величко В.Ю. та ін.] / Патент на корисну модель № 72914. – Опубл. 27.08. 2012, Бюл. № 16.

Получено 24.10.2013