УДК 004.62

С. Н. Калегин, аспирант, ЗАО «МНИТИ» (Московский научно-исследовательский телевизионный институт) (Москва, Россия) (e-mail: skalegin@inbox.ru, ksn@mniti.ru)

РАСПОЗНАВАНИЕ ЯЗЫКА НЕСТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ В СИСТЕМАХ ГЛОБАЛЬНОГО МОНИТОРИНГА

В статье представлена проблема распознавания языка неструктурированных информационных блоков в системах глобального мониторинга, без решения которой невозможно полностью автоматизировать процесс обработки входящих данных, и показаны возможные способы её решения с целью выявления их преимуществ и недостатков, а также выбора универсальной технологии языковой идентификации, подходящей для всех подобных систем, что позволит унифицировать соответствующие программные модули обработки входящей информации.

Современные способы решения указанной проблемы могут быть разделены на 2 категории: символьные и словарные, которые составляют соответствующие идентификационные принципы и обусловливают результаты их применения. Для определения эффективности этих способов автором проведены специальные исследования, позволившие выявить их особенности и определить возможность применения в автоматизированных системах мониторинга. Данные исследования показали, что существующие технологии и программные решения в сфере языковой идентификации информации имеют общие недостатки, которые обусловлены выбором принципов и способов идентификации, каждому из которых присущи характерные преимущества и недостатки. По этой причине разработчикам подобных систем следует обращать внимание на технологическую основу, которой обусловлены характеристики, результаты работы и условия применения конечного продукта. Кроме того, по результатам исследований видно, что универсальных способов или технологий определения языковой принадлежности информационных блоков сегодня не существует по объективным причинам и в каждой конкретной ситуации требуется их подбор под решаемые задачи, что является стимулом для соответствующих специалистов к поиску новых решений.

Ключевые слова: *система глобального мониторинга информации, языковая идентификация информации, способ языковой идентификации, определение языка текста, определение языковой принадлежности информации.*

Ссылка для цитирования: Калегин С. Н. Распознавание языка неструктурированной информации в системах глобального мониторинга // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2017. Т. 7, № 2(23). С. 20–27.

***

В современном информационном мире государственным структурам и коммерческим организациям становится всё труднее следить за динамично меняющимся информационным полем, которое содержит множество различных неструктурированных данных. При этом большая часть информации представлена в текстовом виде (или в него конвертируется), что существенно затрудняет экспресс-обзор по нужной тематике. Поэтому регулярного просмотра новостных лент в СМИ или Интернете и поиска необходимой информации в соответствующих системах не всегда достаточно, так как объём данных, скорость их накопления и разнородность не позволяют человеку провести полный анализ в нужной сфере за приемлемое время [1]. Для этих целей созданы автоматизированные системы глобального мониторинга, позволяющие получать требуемую информацию из различных источников, во всех видах и на разных языках, а также проводить её анализ и создавать нужные отчёты в автоматическом режиме. На основе таких отчётов можно не только анализировать состояние дел в заданной области, но и строить прогнозы развития ситуации, что необходимо для принятия верных решений и стратегического планирования.

На сегодняшний день создано множество систем мониторинга, предназначенных для сбора и обработки определённой информации. Например, система «Астарта» компании Cognitive Technologies, представляющая собой экспертный рубрикатор для сбора, хранения и семантического анализа текстовых ма-

териалов; информационно-аналитическая система «Медиалогия» одноименной компании, позволяющая производить разнообразный анализ данных по определённой тематике; программа TextAnalyst от НПИЦ "МикроСистемы", которая является инструментом смыслового поиска информации и формирования электронных архивов, и многие другие. Последняя способна строить семантические деревья по отдельным статьям, в результате чего создаётся смысловой портрет каждого текста на основе количества упоминаний и частоты встречаемости значимых слов. В TextAnalyst также имеется модуль, генерирующий реферат текстового документа, а многие из подобных систем снабжены внешними или встроенными авторубрикаторами и аннотаторами, которые становятся фактически стандартными атрибутами информационно-аналитических систем. Однако в настоящее время наиболее востребованы системы глобального многоязычного поиска, которые могут собирать информацию на различных языках со всего мира и выдавать унифицированный результатирующий отчёт, а в таком случае возникает проблема языковой идентификации собранных данных. Без решения этой проблемы невозможно полностью автоматизировать процесс обработки входящих информационных блоков, что существенно затормаживает развитие подобных систем.

Для решения указанной проблемы разработано множество способов языковой идентификации, наиболее популярными из которых являются следующие [2]:

использование уникальных символов;

• использование статистики комбинаций символов (n-грамм);

• использование словарей;

• поиск характерных коротких слов.

Эти способы могут быть разделены на 2 категории: символьные и словарные, которые составляют соответствующие идентификационные принципы и обусловливают результаты их применения. Для определения эффективности этих способов автором были проведены специальные исследования, которые позволили выявить их особенности и определить возможность применения в автоматизированных системах мониторинга.

При обработке многоязычной информации в автоматическом режиме перед машиной возникает несколько задач, связанных с выделением блоков на каждом языке и работой с ними. Для их решения создаются специальные программные модули и системы – языковые определители. Место такого определителя в системе глобального мониторинга информации показано на рисунке.

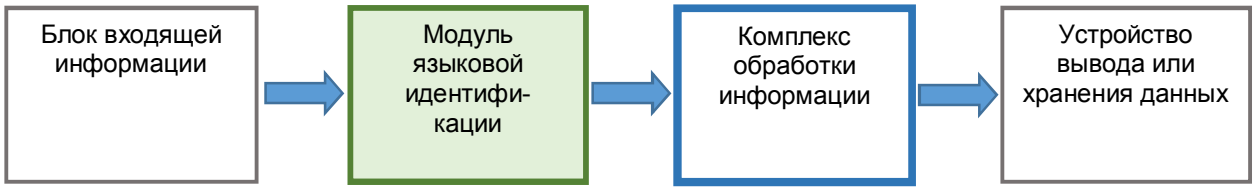

Рис. Место модуля языковой идентификации в системе глобального мониторинга информации

Как следует из рисунка, языковой определитель является первым фильтром и обработчиком входящего информационного блока. Однако каждая подобная программа имеет свои преимущества и недостатки, обусловленные принципом и способом идентификации, заложенными в её основу, или эффективностью корреляции нескольких принципов при их объединении [3], а также их технической реализацией и особенностями результати-

рующего алгоритма, что будет показано далее.

Большинство разработчиков предлагают разделять разноязычные блоки на основании различий письменных систем [4], то есть использовать символьный принцип, что частично рационально для тех языков, традиционные письменности которых существенно отличаются. Например, из текста на английском языке легко выделяются цитаты на арабском, если они записаны арабскими буквами. Однако, как показывает практика, языковая идентификация таких блоков будет носить вероятностный характер и не всегда соответствовать действительности, на что справедливо обращают внимание некоторые специалисты [5, 6]. Например, арабской письменностью фиксируется информация не только на арабском языке, но и на фарси, дари, пушту, кашмири, пенджаби, синдхи, хауса, фула и т.д.; латинский алфавит применяется для записи текстов на сотнях различных языков, включая языки всех континентов, а кириллица используется славянскими, тюркскими, уральскими и другими народами, населяющими Восточную Европу, Кавказ и территорию бывшего СССР.

Ниже приведены фрагменты текстов на языках различных генеалогических групп, использующих латиницу, которые наглядно демонстрируют несостоятельность данного подхода к определению языковой принадлежности информации.

Итальянский язык (романская группа)

Il Talenti propone un nuovo modello per la Cattedrale, che modifica in modo sostanziale il progetto arnolfiano allargando notevolmente l'ampiezza delle campate e riducendo a tre il loro numero in modo da lasciare inalterata la lunghezza delle navate.

Валлийский язык (кельтская группа)

"A dweud y gwir allwn ni ddim ag enwi un peth. Wy'n gwybod y byddai hynnyn beth eitha' ystrydebol i ddweud, ond... mae e i gyd yn deillio mewn gwirionedd ohonof fi'n pigo off beth mae pawb arall yn y band yn cael eu dylanwadu ganddo.

"Mae gan Robin chwaeth hollol wahanol i'r gweddill ohonon ni, ac mae pawb yn codi eu dylanwadau eu hunain ac yn eu rhoi nhw at ei gilydd. Mae'r caneuon i gyd wedi'i gwneud gan un syniad bychan neu ychydig o syniadau bychain.

Голландский язык (германская группа)

Het Mobiele Verkeersplein is door Veilig Verkeer Nederland (VVN) en Shell in het leven geroepen, om kinderen bewuster en veiliger om te leren gaan met het verkeer. Precies vijf jaar geleden werd dit gezamenlijke project op dezelfde locatie in Mijdrecht opgestart. Inmiddels hebben 75.000 kinderen aan het levensechte parcours met verkeersborden, skelters, fietsjes, zebrapaden en verkeerslichten meegedaan. Na menig succesvolle jaren van samenwerking gaat VVN verder met praktisch verkeersonderwijs.

Идо (искусственный язык)

En Nederlando existas granda sabloduni. Sen l' interveno di la naturamiki hike ja de longe esus foresto. En altra distrikto trovesas granda erikeyi. La naturo ne prizas nur unspeca plantaro e probas kreskigar arbori e herbi di fekunda mixuro. Ma se irgaloke videbleskas birketo en erikeyo, naturamiki quik efacas ol.

Суахили (группа банту)

Hadi sasa watu 36 wameripotiwa kuokolewa kutokana na kazi kubwa, iliyofanywa na vikosi vya uokoaji vya Jeshi la Polisi kwa kushirikiana na wananchi wanaoishi katika vijiji vya pembezoni mwa Ziwa Nyasa.

Chanzo cha habari kutoka eneo la tukio, kililiambia gazeti hili jana kuwa boti hiyo ilizama baada ya kugonga mwamba ndani ya ziwa hilo saa tisa alasiri jirani na kilipo kijiji cha Mkenda wilaya ya Nyasa mkoani Ruvuma.

Из приведённых примеров видно, что по знакам письменности определить язык практически невозможно, так как в большинстве случаев она не связана с каким-либо языком непосредственно, а является только одним из возможных средств для фиксации информации, что отражено в определениях понятий «язык» и «письменность».

Однако, несмотря на очевидные недостатки, многие специалисты предлагают использовать особенности традиционной для конкретного языка письменности в целях языковой идентификации текстов на этом языке [7, 8]. В качестве языковых маркеров авторы подобных способов и подходов предлагают использовать характерные символы или сочетания, которые обычно применяются для фиксации информации на данном языке, или статистически наиболее употребительные сочетания символов – n-граммы (табл.) [9], которые могут объединяться в специальные наборы – языковые модели. Такие символы и сочетания называются «идентификационными маркерами» языка, некоторые примеры которых приведены ниже.

Характерные символы различных письменностей

Кириллические: Ё, Й, Ъ, Ь, Ғ, Ҙ, Ҥ, Ц, Ч, Ҟ …

Латинские: Ł, À, Æ, Ê, Ï, Ü, Ä, Ö, Õ, Ñ, Í, Č, Ď, Ħ, Ş, Ŝ, Ž …

Характерные сочетания символов в разных языках

SCH, TSCH, CK, CZ, DZ, DŹ, DŻ, RZ, SZ, SZCZ, ŚĆ, ŹDŹ, ŻDŻ …

Фрагмент таблицы n-грамм для нескольких языков

Шведский	Английский	Немецкий	Французский	Итальянский
en_	_th	en_	_de	_di
_._	he_	er_	es_	to_
er_	the	_de	de_	_de
et_	_,_	der	ent	di_
tt_	nd_	ie_	nt_	_co
_de	ed_	ich	_le	la_
ar_	_an	sch	e_d	re_
-,-	and	ein	le_	ion
fr	_._	che	ion	ent
om_	_to	die	s_d	e_d
_oc	ing	ch_	e_l	le_

Однако такой подход тоже не даёт высокой точности распознавания языка информационных блоков, так как подобные символы и их сочетания могут встречаться в текстах на различных языках. Для примера можно сравнить следующие фрагменты на немецком и турецком языках, содержащие умляуты (буквы с тремой: «**ä**» «**ö**» и «**ü**»), являющиеся отличительной особенностью немецких текстов.

Немецкий язык

„Wir nehmen im Match diesen Ball auf und werden ihn weiter voranspielen“, sagte Schäubles Sprecher in Berlin. Die sogenannten Panama Papers seien keine Überraschung, erhöhten aber den Druck auf Steueroasen auf der ganzen Welt. Das „Unterholz“ bei Versuchen, die Steuerbehörden

auszutricksen, müsse besser ausgeleuchtet werden. Schäuble selbst werde vor der Frühjahrstagung des Internationalen Währungsfonds (IWF) und der Weltbank vom 15. bis zum 17. April in Washington die Initiative ergreifen in der Frage, wie es international mehr Transparenz gegen illegale Finanzgeschäfte geben könne, kündigte Jäger an. (Из новостной ленты)

Турецкий язык

Uzun yıllar önce Çin'de bir kral vardı. Kralın sarayı çok büyük ve çok güzeldi. Çatısı altındı. Pencerelerinde bin tane lâmba vardı. Koridorları uzun ve bahçeleri sayısızdı.

Sarayın çevresinde yeşil bir orman ve mavi bir deniz vardı.

Ormanda sayısız hayvan vardı. Fakat hayvanların en meşhuru küçük gri bülbüldü. Sesi çok güzel ve harikaydi. İnsanlar her yerde bülbülün güzel şarkılarından bahsetti.

Balıkçılar deniz kenarında bülbülün güzel sesini dinledi. Herkes bülbülün güzel şarkılarını duydu ama kimse onu görmedi.

Bülbül Çin'de ve komşu ülkelerde meşhur oldu. Uzak ülkelerden insanlar bülbülü dinlemek için ormana geldiler. Şairler bülbül için şiirler yazdılar. Ülkede herkes bülbülün ününü duydu. Yalnız ülkenin kralı bundan haberdar değildi. (Отрывок из сказки Kral ve bülbül)

Как видно из приведённых фрагментов, по характерным для традиционной письменности конкретного языка символам также невозможно гарантированно идентифицировать язык информационного блока, как и без них, в силу их широкой распространённости. А значит, определять языковую принадлежность текста или речи разумно только по форме их компонентов и лексико-грамматическим элементам, то есть морфемам, словам, синтагмам и т.д.

Для повышения надёжности языковой идентификации разработаны словарные способы, позволяющие определять язык по характерным коротким, служебным или всем значимым словам текста, для чего к программе подключаются специальные списки или словари всех идентифицируемых языков. Однако такой подход к решению проблемы многократно повышает ресурсоёмкость процесса языковой идентификации в целом и не даёт ожидаемых результатов, так как формы слов (особенно коротких) могут полностью совпадать в различных языках, а словарные варианты встречаются далеко не во всех предложениях, что показано на следующих примерах, где в качестве идентификационных маркеров (выделены полужирным шрифтом) используются словарные формы (для русского языка) и короткие слова (для польского).

Русский язык

*Нова была красивой планетой, первой успешной земной колонией. **Сейчас** это **пустыня**. Целые города исчезли с её лица, уничтоженные взрывами нейтронных бомб. Нечего **опознать**. Нечего **похоронить**. Некого **оплакать**. **Вторжение** началось **внезапно**. Объединённые силы Земли нанесли коварные удары по всей территории планеты. **После** них остались растерзанные тела. Крики женщин и детей о помощи захлебнулись в плазменном огне, прожегшем их **плоть**. Мы слышали **много** проповедей о добродетелях прогресса и науки. Что хорошего от них, если целые цивилизации разрушаются в **мгновение** ока?* (Вступление из компьютерной игры "Power DOLLS")

Польский язык

*Przewodniczący Rady Europejskiej opowiadał też, że podczas jego rozmów z innymi premierami **i** prezydentami największe zainteresowanie budziły dwie sprawy. - Byłem trochę tym zaskoczony, ale **po** chwili zrozumiałem dlaczego. **To** były pytania **o** Puszczę Białowieską **i o** stadninę koni. Może dlatego, że **to ma** też wymiar symboliczny. Muszę powiedzieć, że dla mnie jako Polaka ktoś, kto wycina starodrzew al-*

bo doprowadza ***do*** *śmierci koni -* ***i to*** *w takiej stadninie jak Janów - robi straszne rzeczy. Zastanawiam* ***się****, czy nie zaczną też strzelać* ***do*** *bocianów – powiedział.* (Отрывок из газеты)

В первом примере пропускаются целые предложения, а во втором встречаются такие слова, как **to, ma, i, o, do**, которые присущи текстам и на других языках (английском, итальянском, португальском, гаэльском и т.д.). Следовательно, для верной языковой идентификации такие маркеры также малопригодны.

Таким образом, обработка разноязычных информационных блоков требует разработки комплексных алгоритмов, что существенно усложняет технологию языковой идентификации в целом, но повышает её надёжность и результативность. Например, использование в одном определителе сразу нескольких способов языковой идентификации, основанных на различных принципах, существенно повышает вероятность верного определения языка информационного блока за счёт гибридного подхода. Но при этом возникает проблема определения верности результатов применения каждого способа в результатирующем алгоритме, так как адекватной технологии для этого не существует, а рейтинг показателей вероятности не всегда соответствует действительности. Следовательно, простого объединения разных способов в одной программе или системе языковой идентификации недостаточно для гарантированного определения языка, а значит, требуется разработка других технологий, возможно, с привлечения специальных лингвистических знаний. Кроме того, объединение нескольких способов требует привлечения дополнительных ресурсов, например n-грамм-моделей и словарей, а также многократно увеличивает требования к вычислительной мощности аппаратного обеспечения, что может сделать процесс языковой идентификации нерационально ресурсоёмким.

В ходе проведённого исследования были определены следующие критерии, выявляющие основные преимущества и недостатки, а также результативность, ресурсоёмкость и целесообразность применения конкретного способа, технологии, программы или системы:

- основополагающий идентификационный принцип;
- объём и структура идентификационного набора;
- длина (объём) идентификационного маркера;
- количество и качество идентификационных маркеров;
- требуемое количество операций сравнения;
- требуемый объём оперативной памяти;
- требуемая вычислительная мощность компьютера;
- необходимость использования дополнительных систем, моделей, баз данных, словарей и т.п.;
- вероятность верного результата.

Как видно из приведённого списка, на результат машинного определения языковой принадлежности текста тем или иным способом существенное влияние оказывают как принцип и способ идентификации, на базе которого строится логика алгоритма данной технологии, программы или системы, так и технические показатели конкретной реализации.

По итогам проведённых исследований также следует отметить, что все современные технологии языковой идентификации так или иначе зависят от письменной системы, применяемой в анализируемом тексте, а также правил грамматики и орфографии каждого анализируемого языка, что делает невозможным их применение при записи текста нетрадиционным или несовременным способом. Кроме того, для верной идентификации определителю часто требуется существенно больше анализируемого материала, чем декларируют разработчики, что ставит под сомнение заявляемые ими

преимущества рассмотренных способов или программных решений.

Резюмируя сказанное выше, следует отметить, что, несмотря на большое количество и разнообразие существующих технологий и программных решений для определения языковой принадлежности неструктурированного текста, они имеют общие недостатки, которые обусловлены не совершенством алгоритма и не его технической реализацией, а выбором принципов и способов идентификации, каждый из которых имеет характерные особенности [10]. По этой причине при разработке соответствующего программного обеспечения важно не только и не столько стремиться к совершенству алгоритмов и технических решений, сколько обращать внимание на их технологическую основу, которой обусловлены основные характеристики, результаты работы и условия применения конечного программного продукта. Кроме того, результаты исследований показали, что универсальных способов или технологий определения языковой принадлежности информационных блоков сегодня не существует по объективным причинам и в каждой конкретной ситуации требуется их подбор под решаемые задачи. Это является стимулом для разработчиков систем глобального мониторинга к поиску новых решений.

Список литературы

1. Опарин А. Системы мониторинга и анализа СМИ [Электронный ресурс] // PC Week/RE. 2003. № (413) 47. URL: https://www.pcweek.ru/themes/detail.php?ID=66333.

2. Grefenstette G. Comparing two language identification schemes / 3rd International Conference JADT 1995 (Statistical Analysis of Textual Data). Rome, Italy.

3. Калегин С. Н. Важность выбора основного идентификационного принципа при проектировании языковых определителей // Современные информационные технологии и ИТ-образование. 2016. Т. 12, № 2. С. 194–204.

4. Лидовский В.В. Первичная машинная обработка текста: методика и проблематика: монография. М., 1997. Деп. в ИНИОН РАН. 1998, № 53656.

5. Malek Boualem, “Multilingual text editing and Arabic language processing”, AI’95, Conférence en Génie linguistique, Montpellier, France, 1995.

6. Malek Boualem, Jérôme Vinesse Multilingual text processing difficulties, EURESCOM, P923, 1999.

7. Гиляревский Р. С. Определитель языков мира по письменностям. М., 1961. 303 с.

8. Пат. 2251737 Рос. Федерация, G06K9/68. Способ автоматического определения языка распознаваемого текста при многоязычном распознавании / Анисимович К. В., Терещенко В. В., Рыбкин В. Ю.; Аби Софтвер Лтд. (CY); опубл. 10.05.2005.

9. Калегин С. Н. Способы определения языковой принадлежности неструктурированного текста в мультиязычной информационной среде // Международная конференция «CONCORT-2016». Н. Новгород, 2016.

10. Калегин С. Н. Оценка эффективности методов определения языковой принадлежности неструктурированного текста и варианты их программной реализации // Международная конференция «CONCORT-2016». Н. Новгород, 2016.

Поступила в редакцию 13.03.17

UDC 004.62

S. N. Kalegin, Post-Graduate Student, Moscow Scientific Research Television Institute **(**Moscow, Russia) (e-mail: skalegin@inbox.ru, ksn@mniti.ru)

LANGUAGE IDENTIFICATION OF UNSTRUCTURED INFORMATION IN GLOBAL MONITORING SYSTEMS

The article presents the problem of language identification of unstructured information blocks in global monitoring systems, without it's solution is impossible to completely automate the process of incoming data processing, and shows possible methods to solve the problem in order to identify advantages and disadvantages, as well as the choice of a universal language identification technology suitable for all similar systems, which will allow to unify the appropriate software modules for incoming information processing.

Modern solutions of this problem can be divided into two categories: symbolic and vocabulary, which corresponding identification principles and determine the results of their application. To determine the effectiveness of these methods, the author carried out special studies that made it possible to identify their features and determine the possibility of using them in automated monitoring systems. These studies have shown that existing technologies and software solutions in the field of language identification have common shortcomings, which are due to the choice of identification principles and methods, each of which has its own inherent advantages and disadvantages. For this reason, developers of such systems should pay attention to the technological basis, which determines the characteristics, results of work and the conditions for using the final product. In addition, according to the research results, it is clear that there are no universal methods or technologies for language identification of information blocks today for objective reasons and in each specific situation they need to be selected for the tasks to be solved, which is an incentive for the relevant specialists to search for new solutions.

Key words: *system of global information monitoring, language identification of information, language identification method, text language definition, language detection of information.*

For citation: Kalegin S. N. Language identification of unstructured information in global monitoring systems, Proceeding of the Southwest State University. Series: Control, Computer engineering, Information science. Medical instruments engineering, 2017, vol. 7, no. 2(23), pp. 20-27.

***

References

1. Oparin A. Sistemy monitoringa i analiza SMI [Ehlektronnyj resurs] // PC Week/RE. 2003. № (413) 47. URL: https://www.pcweek.ru/themes/detail.php?ID=66333.

2. Grefenstette G. Comparing two language identification schemes / 3rd International Conference JADT 1995 (Statistical Analysis of Textual Data). Rome, Italy.

3. Kalegin S. N. Vazhnost' vybora osnovnogo identifikacionnogo principa pri proektirovanii yazykovyh opredelitelej // Sovremennye informacionnye tekhnologii i IT-obrazovanie. 2016. T. 12, № 2. S. 194–204.

4. Lidovskij V.V. Pervichnaya mashinnaya obrabotka teksta: metodika i problematika: monografiya. M., 1997. Dep. v INION RAN. 1998, № 53656.

5. Malek Boualem, "Multilingual text editing and Arabic language processing", AI'95, Conférence en Génie linguistique, Montpellier, France, 1995.

6. Malek Boualem, Jérôme Vinesse Multilingual text processing difficulties, EURESCOM, P923, 1999.

7. Gilyarevskij R. S. Opredelitel' yazykov mira po pis'mennostyam. M., 1961. 303 s.

8. Pat. 2251737 Ros. Federaciya, G06K9/68. Sposob avtomaticheskogo opredeleniya yazyka raspoznavaemogo teksta pri mnogoyazychnom raspoznavanii / Anisimovich K. V., Tereshchenko V. V., Rybkin V. Yu.; Abi Softver Ltd. (CY); opubl. 10.05.2005.

9. Kalegin S. N. Sposoby opredeleniya yazykovoj prinadlezhnosti nestrukturirovannogo teksta v mul'tiyazychnoj informacionnoj srede // Mezhdunarodnaya konferenciya «CONCORT-2016». N. Novgorod, 2016.

10. Kalegin S. N. Ocenka ehffektivnosti metodov opredeleniya yazykovoj prinadlezhnosti nestrukturirovannogo teksta i varianty ih programmnoj realizacii // Mezhdunarodnaya konferenciya «CONCORT-2016». N. Novgorod, 2016.