Калегин Сергей Николаевич

СПОСОБЫ ОПРЕДЕЛЕНИЯ ЯЗЫКА ТЕКСТА

Цель данной статьи - представить современное состояние проблемы идентификации языка текста в виде обзора известных способов её решения с указанием их преимуществ и недостатков. Большинство этих способов могут использоваться как с применением компьютеров (машинной обработки), так и без них. Предлагаемый обзор наглядно показывает сильные и слабые стороны каждого метода с указанием условий его использования. Кроме того, в работе сделан акцент на математические способы определения языковой принадлежности текста. В завершении статьи автор предлагает свой вариант языковой идентификации текста.

Адрес статьи: www.gramota.net/materials/2/2015/12-2/21.html

Источник

Филологические науки. Вопросы теории и практики

Тамбов: Грамота, 2015. № 12(54): в 4-х ч. Ч. II. С. 84-89. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2015/12-2/

© Издательство "Грамота"

Информация о возможности публикации статей в журнале размещена на Интернет сайте издательства: www.gramota.net

Вопросы, связанные с публикациями научных материалов, редакция просит направлять на адрес: phil@gramota.net

УДК 81-139
Филологические науки

Цель данной статьи – представить современное состояние проблемы идентификации языка текста в виде обзора известных способов её решения с указанием их преимуществ и недостатков. Большинство этих способов могут использоваться как с применением компьютеров (машинной обработки), так и без них. Предлагаемый обзор наглядно показывает сильные и слабые стороны каждого метода с указанием условий его использования. Кроме того, в работе сделан акцент на математические способы определения языковой принадлежности текста. В завершении статьи автор предлагает свой вариант языковой идентификации текста.

Ключевые слова и фразы: способ определения языка; языковая идентификация текста; машинная обработка текста; определение языковой группы текста; языковая принадлежность текста.

Калегин Сергей Николаевич
Московский научно-исследовательский телевизионный институт
ksn@mniti.ru

СПОСОБЫ ОПРЕДЕЛЕНИЯ ЯЗЫКА ТЕКСТА©

В данной статье речь пойдёт о методах (способах) определения языка текста применительно к вычислительной технике (компьютерам), что позволяет отнести работу к сфере компьютерной лингвистики. Представленный материал будет интересен как специалистам в области языковедения, так и техническим специалистам, решающим задачи в области машинной обработки текста. Ниже даётся обзор популярных способов машинной идентификации языка текста, полученного компьютером из внешних источников. Большинство этих способов являются универсальными, т.к. могут использоваться и без применения компьютеров, но сегодня такой подход встречается довольно редко и практически не востребован. По этим причинам в дальнейшем изложении будет подразумеваться обработка текста с помощью вычислительной техники, если явно не указано иное.

Также нужно отметить, что в данной статье не рассматривается проблема различных кодировок символов и вопросы их взаимных преобразований, так как эта тема не имеет прямого отношения к процессу идентификации языка текста и требует отдельного исследования в силу её сложности и большого количества сопутствующего материала. По тем же причинам в статье не затрагиваются вопросы идентификации речи.

Для лучшего восприятия изложенного ниже материала будет рационально и целесообразно вначале привести определения основных понятий (*даны автором статьи – С. К.*), на которых будут базироваться все размышления, утверждения и выводы, ввиду неоднозначности и «размытости» их интерпретации, а также множества толкований и определений, которые встречаются в различных словарях. Это поможет избежать недоразумений и внесёт в изложение некоторую конкретику.

Текст – это набор слов, словосочетаний, синтагм или предложений, представляющих собой семантическое единство.

Язык – это лексико-грамматическая система, выполняющая коммуникативную функцию.

Письменность (письменная система) – это законченная система обусловленных знаков, предназначенная для фиксации информации.

Машинная (компьютерная) обработка – это комплекс целенаправленных действий, осуществляемых машиной (компьютером) для выполнения определённой задачи.

Проблема машинного определения языка текста сегодня актуальна как никогда раньше вследствие развития электронно-вычислительных машин и их популярности в современном мире. Машины помогают людям многократно ускорить поиск, передачу и обработку информации, а также обмениваться опытом, публиковать свои идеи и обсуждать чужие. Информационные сети и мобильные компьютеры стали частью жизни человека и он, во многих случаях, старается переложить на них многочисленные рутинные дела и процессы, которые требуют больших энергетических и временных затрат. И с каждым годом эта тенденция растёт. Именно поэтому любые разработки в области компьютерной техники являются актуальными и востребованными в современном обществе. Кроме того, с глобализацией и развитием коммуникаций возрастает необходимость международного общения, что ведёт к потребности межъязыковых переводов текстов. Последнее обстоятельство вынуждает людей искать или создавать средства и способы упрощения работы со множеством языков. Например, для облегчения международной коммуникации были созданы специальные программы и устройства-переводчики, которые призваны помочь человеку в работе с текстами на различных языках. Кроме того, на сегодняшний день создано множество различных электронных каталогов, библиотек, сетевых баз данных и других подобных информационных систем, которые содержат и обрабатывают информацию на десятках и сотнях различных языков, что приводит к необходимости их автоматической (или полуавтоматической) идентификации уже на этапе получения информации данной системой. Без такой идентификации было бы невозможно корректно распределить информацию в базе данных, а также определить какие модули

© Калегин С. Н., 2015

потребуются для её обработки или какому специалисту она должна быть направлена. Более того, необходимость определения языка возникает и у простого офисного сотрудника, например, при работе с корреспонденцией или поиске информации по нужной теме, а также у программистов, которые пытаются автоматизировать процессы ввода и обработки текстов. К примеру, при проверке орфографии и грамматики вводимого в машину текста обязательно нужно знать, на каком языке этот текст написан, чтобы выбрать для него подходящий словарь или справочник. По этим причинам люди начали задумываться о способах определения языка заданного текста и о реализации этих способов в алгоритмах прикладных программ, которые подключаются в виде модулей к системам обработки информации, браузерам и текстовым процессорам. Ниже приводится краткое изложение сути наиболее популярных методов языковой идентификации.

Использование словарей. Этот способ заключается в переборе словарей множества языков и поиске совпадений слов текста со словами в данных словарях. Метод относительно простой и легко реализуемый как программными средствами, так и без них. Однако энергетические и временные затраты на поиск каждого слова текста в каждом словаре множества языков делают данный способ практически нецелесообразным из-за большого количества операций. К тому же, при простом сравнении слов текста со словарными формами их грамматические вариации не учитываются, а значит, вероятность обнаружения совпадений слов со словами одного языка резко понижается (особенно в небольших текстах), что приводит к ошибкам идентификации. Более того, при использовании данного способа требуется располагать словарями всех идентифицируемых языков и их нужно где-то хранить, что требует выделения дополнительных ресурсов.

Учитывая специфику данного метода, его можно использовать только при заранее известном небольшом количестве языков. А при машинном определении языка текста потребуется достаточно мощный и дорогостоящий компьютер или же этот процесс займёт много времени, что может сделать саму идентификацию нецелесообразной.

Использование уникальных знаков. Данный способ заключается в отличии языков друг от друга по особенным буквам (или знакам), в частности, по буквам с диакритическими значками (диакритиками), которые используются при записи текстов на данном языке. Диакритика бывает над гласной (например, в буквах «й» или «ё»), над согласной (буква «č») или может как-то иначе сопровождать букву (или буквосочетание). Кроме того, во многих алфавитах дополнительно используются специальные символы для обозначения фонем (звуков) данного языка. Например, в польском алфавите есть знак, напоминающий перечёркнутую букву «L», а в украинском – буква, похожая на русское «Э», но повёрнутая в другую сторону. Плюс ко всему, некоторые языки имеют собственную письменную систему, как например, японский или корейский. Это наталкивает многих программистов на ассоциацию конкретной письменности с определённым языком, что не всегда приводит к ожидаемому результату. Например, если определять русский язык по наличию в тексте буквы «ё», то многие технические тексты, где данная буква практически не используется, не будут идентифицированы, а вот тексты на других языках (допустим тюркских), записанных кириллицей и имеющих подобную фонему (звук), как раз будут отнесены к русскому. Подобным образом язык часто определяют системы распознавания текста. Например, в описании «Способа автоматического определения языка распознаваемого текста при многоязычном распознавании» приводится следующее:

…предположительно содержащие признаки изображения символов текста, с последующим сопоставлением изображения в блоках с эталонным изображением, в нескольких специальных признаковых (или растровых) классификаторах, содержащих символы одного определенного языка [1, с. 8]…

и далее в том же документе:

Вместо нескольких отдельных классификаторов иногда используют единственный, содержащий признаки символов всех языков, предположительно присутствующих в документе [Там же, с. 9].

Такой способ представлен, например, в патенте США № 6370269 April 9, 2002 [4].

Из приведённых цитат следует, что авторы данных способов (изобретений) неразрывно связывают распознаваемые символы (буквы, слоговые знаки или иероглифы) с определёнными языками, что с точки зрения лингвистики, по мнению автора статьи, в корне ошибочно.

Таким образом, недостатком данного способа является смешение понятий языка и письменности, что приводит к грубейшим ошибкам идентификации. Как следует из определений, данных вначале статьи, практически все письменности подходят для фиксации мыслей или образов (информации), выражаемых в словах и синтагмах, а значит, их можно использовать для записи текста на любом языке. Например, по-русски можно писать как кириллицей, так и латиницей (а также еврейскими, арабскими или греческими буквами), и в любом направлении, что не приведёт к каким-либо искажениям передаваемой информации. Это подтверждается многочисленными примерами народов, которые с лёгкостью переходили с одной письменной системы на другую за короткое время, и это никак не отражалось на их языке. Для примера можно взять языки средней Азии и Кавказа, на которых писали различными письменностями в различные исторические периоды, а сейчас они используют модифицированную кириллицу или латиницу. Более того, тексты на некоторых языках могут быть записаны несколькими письменными системами без особой разницы. Например, на языках бывшей Югославии и сейчас пишут либо латиницей, либо кириллицей, а на вьетнамском языке чаще всего пишут латинскими буквами с диакритиками, хотя существует собственная вьетнамская письменность, созданная на основе китайских иероглифов (которые когда-то также использовались для записи вьетнамских текстов). Кстати говоря, любые иероглифы представляют собой упрощённые рисунки (пиктограммы), которые фиксируют не звучание слов,

а идею (мысль) или образ, поэтому могут служить для записи текстов на абсолютно любом языке, что подтверждается наличием подобных знаков в древности у различных народов и их заимствованием друг у друга.

Использование статистики комбинаций символов (байтовых последовательностей [2] **или n-грамм).** Данный способ обычно заключается в определении языка по количеству типичных комбинаций символов, характерных для конкретного языка. Чаще всего подсчитываются комбинации двух (диграммы) или трёх (триграммы) символов, хотя могут встречаться и другие варианты n-грамм. Таким образом, текст ассоциируется с тем языком, которому с наибольшей вероятностью соответствует большинство найденных в тексте комбинаций символов (или байтовых последовательностей). То есть данный способ можно назвать чисто математическим, так как анализ самих символов, слогов или слов не производится. Также не производится транскрипция или транслитерация, а это значит, что уже на стадии формирования n-грамм будет выбран неверный метод деления текста, что однозначно отразится на результате идентификации.

Недостатками данного способа являются его абстрактность и вероятностный результат, так как никакого лексического или грамматического анализа текста не производится, а в завершении процесса определения выдаётся список различных языков (часто даже неродственных), к которым можно было бы отнести данный текст с некоторой вероятностью. Например, при идентификации текста на немецком языке данным способом, он может быть отнесён примерно с одинаковой вероятностью к шведскому и суахили (один из африканских языков), которые не имеют между собой ничего общего. Более того, для приемлемой работоспособности данного метода требуется набрать определённую статистику встречаемости комбинаций символов (или n-грамм) в различных языках, а для этого нужно проанализировать десятки или сотни тысяч текстов и создать солидную базу данных! В силу указанных причин данный способ не очень популярен среди лингвистов, однако, пользуется успехом у математиков и программистов, так как не требует специальных знаний в области языковедения.

Грамматический анализ текста [3]. Суть этого способа заключается в морфологическом разборе слов и синтаксическом анализе предложений. Сама идея очень привлекательна своим естеством. То есть, примерно такой же анализ текста производит и человек при попытке идентификации языка (вкупе с лексическим сопоставлением), что придаёт данному способу определённую натуральность. Однако чтобы провести такой анализ требуются специальные лингвистические модели для каждого определяемого языка (а для большинства языков их просто не существует!) и множество действий с каждым словом текста, что выливается в миллионы операций, на реализацию которых требуется нецелесообразно много ресурсов. Таким образом, несмотря на свою естественность и научный подход, данный способ может использоваться в ограниченных условиях и только для некоторых языков.

Разумеется, все вышеперечисленные способы определения языка текста имеют множество вариаций и комбинаций, позволяющих, так или иначе, улучшить результаты их применения. Однако следует отметить, что машинное определение языка рассмотренными способами является принципиально вероятностным, условно применимыми или нецелесообразно ресурсоёмкими в силу указанных недостатков. А это сильно ограничивает их использование, так как во многих случаях такие результаты неприемлемы. Например, если при наборе текста в текстовом процессоре язык будет определён неверно, то соответственно будет выбран и словарь для проверки орфографии, что повлечёт за собой тотальные ошибки. То же самое произойдёт при неверном определении языка текста, вставленного в программу-переводчик, которая не сможет подобрать нужные словари и грамматические модели для перевода. И в том, и в другом случае работа программы по идентификации языка будет бесполезной, а вероятностный результат в приведённых примерах абсолютно недопустим, как и затрата большого количества ресурсов, которая вызовет «подвисание» (существенную задержку) при выполнении программы.

Поиск служебных слов. Этот способ довольно редко используется, но предлагается программистами с завидной регулярностью. Его основная идея заключается в выделении из текста характерных служебных слов и частиц, таких как союзы, предлоги или артикли. Например, при идентификации английского языка предлагается искать артикль «the». Разумеется, такой подход обычно используется далёкими от лингвистики людьми, а для языковеда обречённость подобного метода идентификации очевидна. Данный способ не учитывает множества совпадений служебных слов в родственных языках и похожих коротких слов, междометий и грамматических форм в других языках. Для примера можно привести романские языки, в большинстве из которых встречается артикль «la» и форма «ma» или «mas». Кроме того, такие же формы встречаются и в других языках, например, в славянских, а также в греческом, эсперанто, идо и т.д. Следовательно, в результате идентификации языка текста описанным способом будет выдан список различных языков, в которых встречаются заданные служебные слова. Такой результат в большинстве случаев является практически бесполезным.

Отдельно следует остановиться на тех случаях, когда в тексте смешиваются слова, записанные разными алфавитами или письменными системами. Например, имена или названия компаний и товаров могут быть написаны на оригинальном языке, а всё предложение сформулировано по-русски. Или же в тексте могут встретиться цитаты на других языках. К примеру, в художественных произведениях наших классиков часто используются фразы и «крылатые» выражения на латинском или французском языке. Также, в связи с развитием компьютерных сетей и сетевого общения, стоит упомянуть о современной тенденции писать текст не традиционной письменностью, которая обычно применяется для данного языка, а использовать наиболее доступные пишущему или наиболее понятные целевой аудитории символы (например, на форумах часто пишут по-арабски кириллицей, изменяя, при этом, направление письма), что никак не предусмотрено упомянутыми способами определения языка. То есть при транскрипции или транслитерации текста, для упомянутых способов он становится неопределяемым в силу специфики их подходов к идентификации.

Итак. Из всего вышеизложенного можно сделать следующие выводы:

1) на сегодняшний день имеется несколько популярных способов машинной идентификации (определения) языка текста, каждый из которых имеет свои недостатки и может быть использован при определённых условиях. Однако универсального способа не существует;

2) ни один из популярных способов не учитывает вероятность транскрипции или транслитерации текста, а также изменение направления письма;

3) все перечисленные способы дают вероятностный результат и требуют затраты значительного количества ресурсов, что не всегда приемлемо и целесообразно;

4) большинство авторов современных способов не использует лингвистические познания, что вызывает смешение понятий и неверные ассоциации (например, языка и письменности), приводя к отрицательным результатам;

5) ни один из упомянутых способов, как правило, не определяет языки цитат и названий, встречающихся в тексте на выбранном языке. То есть, данные способы практически не рассчитаны на многоязычные тексты или же определяют языки с некоторой вероятностью, которая не всегда приемлема в результатах идентификации;

6) в популярных сегодня способах не рассматривается возможность определения языковой группы, что могло бы существенно уменьшить проблему определения языка текста, а в некоторых случаях даже её решить. Для примера можно привести ситуацию с сортировкой текстов в электронных каталогах, бюро переводов, почтовых программах, библиотеках или системах обработки информации, где разница между близкородственными языками не всегда имеет значение.

Таким образом, несмотря на значительные достижения в области лингвистики и автоматизации обработки текста, современные методы машинной идентификации языков далеки от совершенства и здесь ещё есть над чем работать. Эффективность используемых способов недостаточно высока вследствие их принципов идентификации, а также необходимости перебора отдельных слов (и / или словосочетаний) по словарям, создания лингвистических моделей и баз данных, сравнения символов национальных письменных систем или групп символов (байтовых последовательностей) по набранной статистике встречаемости их комбинаций, что при широком спектре идентифицируемых языков требует затраты нецелесообразного количества ресурсов.

Для полноты обзора и расширения сферы применения машинной обработки текстов ниже приводится способ определения языка, который, по мнению автора, обладает значительными преимуществами, как техническими, так и лингвистическими, что позволяет выделить его из списка известных подходов и предложить в качестве альтернативы для реализации в программах идентификации языка текста.

Способ автоматизированного определения языка или языковой группы текста

Предлагаемый способ позволяет определить язык анализируемого текста или языковую группу, к которой он относится. Данным способом могут быть идентифицированы как естественные языки (такие как русский, немецкий, английский, кастильский, латинский и т.д.), так и созданные искусственно (как например: волапюк, эсперанто, идо, интерлингва и т.д.). Суть этого метода заключается в использовании наиболее употребительных глаголов в качестве ключевых элементов идентифицирующей матрицы (фильтра), через которую пропускается текст. А значит, таким образом может быть идентифицирован любой язык, основным связующим элементом или основной частью речи которого является глагол.

Техническим результатом использования предлагаемого способа в компьютерных программах является значительное расширение сферы применения машинной идентификации при улучшении определения языка текста и возможность определения языковой группы в тех случаях, когда язык идентифицировать не удаётся.

Для применения данного метода достаточно составить набор (матрицу) ключевых форм нескольких часто используемых глаголов каждого идентифицируемого языка или языковой группы. То есть, в этом наборе каждый язык или языковая группа соотносятся с грамматическими формами нескольких глаголов и / или их семантически значимыми частями, такими как корни или основы. В качестве идентифицирующих слов используются наиболее употребительные (как например, «делать», «ходить» и т.п.), вспомогательные («быть», «иметь» и т.п.) или модальные (такие как «хотеть», «мочь» и т.п.) глаголы, а для сокращения количества идентифицирующих элементов должны учитываться только наиболее распространённые грамматические формы. Для большинства языков, кроме основных форм глагола, достаточно указать формы настоящего и простого прошедшего времён в действительном залоге изъявительного наклонения, так как в подавляющем большинстве текстов используются именно они. Выбор конкретных форм глаголов зависит от языка, цели и уровня идентификации. Например, для определения только языковой группы и для определения конкретного варианта или диалекта будут использоваться различные наборы глагольных форм. С помощью комбинаций грамматических групп и форм глаголов идентифицируемых языков, и при условии исключения из составляемых наборов совпадающих форм в разных языках и / или языковых группах, может быть достигнута высокая точность идентификации языка или языковой группы текста. От качества составления таких наборов зависит эффективность и область применения описываемого способа, количество идентифицируемых языков и точность определения языковой принадлежности текста.

Данный набор может представлять собой, например, список с определённой структурой, таблицу или многомерный массив, где представлены одна или несколько групп глаголов каждого идентифицируемого языка, указана связь этих групп с конкретным языком или языковой группой (и / или подгруппой), а также языковой ветвью, семьёй или макросемьёй по мере необходимости. Такая иерархия набора идентифицирующих элементов позволяет определять языковые ветви, группы или подгруппы без определения самого языка

анализируемого текста. При этом данная иерархия может быть разветвлённая и многоуровневая (где, например, глаголы близкородственных языков находятся на одном уровне отдельной ветви иерархии), а для каждой языковой группы и каждого языка могут даваться уточнения или более подробная языковая классификация, например, деление на подгруппы, варианты и / или диалекты. К примеру, английский язык относится к германской языковой группе и для него существуют британский, американский и австралийский варианты со множеством диалектов внутри каждого из них.

Составление таких идентификационных наборов глагольных форм с указанием на соответствие конкретному языку или языковой группе (а также с другими нужными индикаторами) является необходимым и единственным достаточным условием для использования данного способа. Эти наборы могут быть составлены как вручную, так и с помощью компьютера в автоматизированном режиме. Более того, составить подобный идентификационный набор для нескольких десятков языков может всего один специалист-языковед за несколько часов работы.

В целом, предлагаемый способ имеет более широкую сферу возможных применений, обеспечивает получение новых технических результатов, обладает рядом преимуществ перед популярными методами, решающими аналогичные задачи, и является одним из наиболее рациональных способов определения языковой принадлежности текста на текущий момент времени. Применение данного способа на практике позволит существенно повысить качество и / или скорость определения языка текста, а технические результаты такого применения позволят существенно сократить затраты и сэкономить время в процессе языковой идентификации.

При реализации данного способа на компьютере значительно сокращается занимаемое программой (и её компонентами) место в оперативной памяти и на устройстве хранения информации, а также потребление вычислительных ресурсов. При этом не требуется использование словарей, грамматических справочников, лингвистических моделей (или графов), баз данных, статистики встречаемости определённых последовательностей символов и т.д. для каждого идентифицируемого языка. Это позволит отводить на идентификацию языка гораздо меньше машинного времени, а также освободить часть ресурсов для решения других задач или создавать менее мощные машины (или менее требовательные к ресурсам программы). Особенно это важно в области *web*-приложений и мобильных компьютеров, которые в последнее время стали неотъемлемым атрибутом повседневной жизни для бóльшей части цивилизованного мира.

Таким образом, данный способ является более универсальным, эффективным и технологичным по сравнению с упомянутыми выше и позволяет значительно улучшить результат по ряду показателей, определить языковую группу, а также упростить и ускорить процедуру идентификации языка текста. Среди преимуществ этого метода можно выделить следующие:

1) возможность работы с многоязычными текстами и точного определения всех языков, используемых в анализируемом тексте, при наличии в нём форм глаголов из идентификационного списка;

2) возможность точного определения языковой семьи, ветви или группы языков, к которой относится язык анализируемого текста (например: славянская, германская, романская, кельтская и т.д.);

3) возможность идентифицировать язык по грамматическим формам и / или их семантически значимым частям (основам или корням) небольшой группы глаголов, например, вспомогательных, модальных, наиболее употребительных и т.д. (в каждой группе по несколько глаголов), или комбинации таких групп;

4) независимость от системы письма или представления информации в анализируемом тексте;

5) значительное повышение точности идентификации языка при небольших объёмах текста;

6) при использовании компьютера имеется возможность обойтись без сложных алгоритмов и мощных вычислительных средств;

7) возможность регулирования функциональности, точности определения и скорости работы с помощью расширения и уточнения или сокращения и упрощения предварительно составляемых наборов форм глаголов идентифицируемых языков или языковых групп;

8) не требуется использования словарей определяемых языков и баз данных, а также изучения грамматики, создания дерева (или модели) грамматических зависимостей, сбора статистики по использованию комбинаций символов и т.д., что позволяет значительно сократить количество выделяемых на обработку ресурсов;

9) текст может быть представлен в любой воспринимаемой компьютером или человеком форме (например, в виде изображений символов, комбинаций точек шрифта Брайля и т.д. с применением одной из известных письменных систем, а также передан в виде блока (набора) сигналов, например, звуковых волн, азбуки Морзе и т.п.), что делает предлагаемый способ более универсальным;

10) количество идентифицирующих элементов и операций сравнения при реализации данного метода в сотни раз меньше, чем при использовании словарей, лингвистических моделей или последовательностей символов (байтовых последовательностей) популярными сегодня способами.

Несмотря на реальные преимущества, этот способ также имеет свои недостатки. Например, очевидно, что работать он будет только при условии наличия в тексте форм глаголов и для отдельных слов (названий, терминов и т.п.) он практически не применим, так как не использует словари для идентификации языка. Однако если подходить к этому формально, отдельное слово текстом не является и по одному слову точно определить язык практически невозможно, ввиду совпадения форм слов в различных языках. По этой причине названный недостаток не умаляет достоинства данного подхода. Зато этот способ даёт явные преимущества при его внедрении и позволяет существенно улучшить результаты языковой идентификации, а также значительно расширить сферу автоматизированной обработки литературы, сократить затраты на такую обработку и ускорить процессы, так или иначе связанные с определением языковой принадлежности текстов.

Заключение

Представленный обзор способов определения языковой принадлежности текста не является исчерпывающим, так как в нём не охвачены многочисленные вариации и комбинации рассмотренных методов, а также более редкие и практически неиспользуемые сегодня подходы к идентификации языка. Но даже краткое и поверхностное изложение затронутой темы показывает множество проблем в данной области, которые ждут своего решения. Это является определённым стимулом для специалистов-языковедов к проведению дальнейших исследований, а для программистов причиной поиска новых оригинальных решений при разработке алгоритмов программ.

Предложенный автором статьи способ языковой идентификации позволяет по-новому посмотреть на рассмотренную проблему и открывает сразу два направления дальнейших изысканий: 1) выявление ключевых слов в каждом языке, однозначно его идентифицирующих; 2) сужение спектра подходящих языков при идентификации до группы или подгруппы, а также выявление определяющих элементов для каждой из них. Решение этих задач позволит значительно продвинуться в решении проблемы языковой идентификации вообще и текста в частности.

Список литературы

1. **Анисимович К. В., Терещенко В. В., Рыбкин В. Ю., Аби Софтвер.** Способ автоматического определения языка распознаваемого текста при многоязычном распознавании: патент № 2251737 РФ, G06K9/68 / Лтд. (CY). Опубл. 10.05.2005.
2. **Лапшин В. А., Пшехотская Е. А., Перов Д. В.** Способ автоматизированного определения языка и (или) кодировки текстового документа: патент № 2500024 РФ, G06F17/00 / «Центр Инноваций Натальи Касперской» (RU). Опубл. 27.11.2013.
3. **Селезнев К.** Обработка текстов на естественном языке [Электронный ресурс] // Открытые системы. 2003. № 12. URL: http://www.osp.ru/os/2003/12/183694/ (дата обращения: 31.10.2015).
4. **Al-Karmi, Abdel Naser, Shamsher S., Baldev Singh.** Optical character recognition of handwritten or cursive text in multiple languages (Оптическое распознавание символов рукописного или курсивного многоязычного текста): патент № 6370269 США / International Business Machines Corporation (USA). Опубл. 09.04.2002.

THE WAYS OF IDENTIFICATION OF TEXT LANGUAGE

Kalegin Sergei Nikolaevich
Moscow Scientific Research Television Institute
ksn@mniti.ru

The article aims at presenting the current state of the problem of identification of the text language in the form of the review of the known ways of its solutions with the indication of their advantages and disadvantages. Most of these ways can be used either with computers (machine processing) or without them. This review shows clearly the strengths and weaknesses of each method indicating the conditions of its use. Besides, the emphasis is put on the mathematical ways for identifying the linguistic belonging of the text. In conclusion the author proposes his own version of the linguistic identification of the text.

Key words and phrases: way of language identification; linguistic identification of the text; machine processing of the text; identification of linguistic group of the text; linguistic belonging of the text.

УДК 8; 80
Филологические науки

Статья посвящена категории неопределенности имен существительных в рамках референциального подхода на материале языка современных русских газетных текстов. Авторы преследуют цель доказать эффективность и особую выразительность разноуровневых средств выражения семантики неопределенности в синтаксическом аспекте в разных жанрах современных российских газет.

Ключевые слова и фразы: семантика неопределенности; референциальный статус; словосочетание и предложение; морфолого-синтаксические показатели; контекст; придаточные предложения.

Кацитадзе Инна Мангуровна, к. филол. н., доцент
Христианова Наталья Валерьевна, к. филол. н.
Южный федеральный университет
mangurowna@yandex.ru; nkhr75@mail.ru

СИНТАКСИЧЕСКИЙ АСПЕКТ КАТЕГОРИИ НЕОПРЕДЕЛЕННОСТИ СУЩЕСТВИТЕЛЬНЫХ В СОВРЕМЕННЫХ РУССКИХ ПЕЧАТНЫХ СМИ©

Синтаксичность категории неопределенности проявляется в пределах таких языковых единиц, как словосочетание и предложение. Важным в дефиниции словосочетания должно быть указание на то, что оно является звеном структуры предложения, и что в нем выражается отношение между двумя мыслительными коррелятами реальных предметов и признаков. На газетной полосе категория неопределенности занимает весомое место.

© Кацитадзе И. М., Христианова Н. В., 2015