Industrial engineering. Management engineering
Информатика, вычислительная техника и управление

http://www.moderntechno.de/index.php/meit/article/view/ge4-005
DOI: 10.30890/2567-5273.2018-04-01-005

**УДК 004.2**

# AUTOMATIC CHECKING OF PUNCTUATION IN RUSSIAN LANGUAGE TEXTS: ELECTRONIC IDIOMS DICTIONARY

# АВТОМАТИЧЕСКАЯ ПРОВЕРКА ПУНКТУАЦИИ В РУССКОЯЗЫЧНЫХ ТЕКСТАХ: ЭЛЕКТРОННЫЙ СЛОВАРЬ ФРАЗЕОЛОГИЧЕСКИХ ОБОРОТОВ

**Polyakova I.N. / Полякова И.Н.**
*Candidate of physico-mathematical Sciences. / к.ф.-м.н.*
*SPIN: 3817 -5512*
**Akhmetova G. R. / Ахметова Г.Р.**
*Moscow State University of Lomonosov, Leninskie gori, GSP-1, Moscow, Russian Federation*
*Московский государственный университет им. М.В.Ломоносова, Москва, Ленинские горы, д.1*

***Аннотация:*** *В статье рассматривается проблема автоматической проверки пунктуации в предложениях русского языка с фразеологическими оборотами. Для этого разработана структура и проведена формализация лингвистического фразеологического словаря русского языка. Электронный словарь представлен xml-файлом, и для удобства его использования написан программный модуль на языке python 3.x. Данный словарь можно применять и в других системах автоматической обработки текста.*

***Ключевые слова:*** *системы автоматической обработки текста, электронные словари, автоматическая проверка пунктуации, формализованный словарь фразеологических оборотов.*

**Вступление.** Письменная речь без знаков препинания или при неточном, неполном, неправильном их применении часто трудна для понимания. Пунктуационная грамотность необходима не только как показатель знания русского языка, общей культуры человека, она играет важную роль в социальном плане.

В данной статье нас будут интересовать правила, в которых используются фразеологизмы. Фразеологизмами или фразеологическими оборотами называются устойчивые по составу и структуре, лексически не делимые по значению словосочетания и предложения, выполняющие функцию отдельной словарной единицы. [1] Для проверки некоторых правил достаточно результатов морфологического и синтаксического анализа предложения, а для остальных необходим семантический анализ. Для правил первой группы можно разрабатывать алгоритмы проверки пунктуации на основе словарей.

**Электронный словарь фразеологических оборотов.** В русском языке для фразеологических оборотов существует лингвистический словарь [1], в котором даны не только сами фразеологизмы русского языка с их толкованиями, но также примеры их стилистического употребления. Большое внимание в словаре уделяется синтаксической роли фразеологизмов, закрепившихся в языке в функции того или иного члена предложения. Помимо этого показана эмоционально-экспрессивная окраска представленных фразеологизмов и стили русского языка, в которых они употребляются.

Пример словарной статьи: «ни рыба ни мясо. *Сказ. Разг., чаще неодобр.*

Заурядная, посредственная личность.»

Применять данный словарь в существующем виде для автоматической обработки текстов невозможно, поэтому необходимо создать электронную версию словаря фразеологических оборотов.

Для каждой словарной статьи в электронной версии словаря будем указывать леммы фразеологизма – нормальные формы слов (именительный падеж для существительных, инфинитив для глаголов и т. д.), его значения, флаг изменяемости, возможные синтаксические роли, а также эмоционально-экспрессивную окраску, если есть соответствующие обозначения в лингвистическом словаре.

Фразеологизмы могут быть неизменяемыми (в этом случае в лингвистическом словаре они представлены в своей единственной форме) или могут частично изменяться. Для указания этого свойства необходим флаг изменяемости.

Возможные синтаксические роли фразеологизма ([1], с. 6): 1 – подлежащее; 2 – сказуемое; 3 – обстоятельство; 4 – определение; 5 – обращение; 6 – вводное слово.

Различны эмоционально–экспрессивные окраски фразеологических оборотов ([1], с. 6): 1 – грубое; 2 – ироничное; 3 – книжное; 4 – неодобрительное; 5 – одобрительное; 6 – отвлеченное; 7 – презрительное; 8 – пренебрежительное; 9 – просторечное; 10 – профессиональное; 11 – разговорное; 12 – устаревшее; 13 – фольклорное; 14 – шутливое; 15 – экспрессивное. Для некоторых фразеологических оборотов в словаре нет определенных синтаксических ролей и эмоционально–экспрессивных окрасок.

Для рассматриваемой ранее лингвистической словарной статьи фразеологизма *«ни рыба ни мясо»* электронная словарная статья примет вид: *«[ни, рыба, ни, мясо] : [«Заурядная личность», «Посредственная личность»], 0, [2], [4, 11]»*. Сначала – список лемм, затем – возможные значения фразеологического оборота, далее значение флага изменяемости (0 – фразеологизм неизменяем), 2 – синтаксическая роль сказуемого, 4 и 11 – разговорная неодобрительная эмоциональная окраска.

**Формализация статей словаря.** Введем формальное описание статей электронной версии словаря фразеологических оборотов:

***<словарная статья>*** **::=** *<леммы фразеологизма> : <значения>, <флаг изменяемости>, <синтаксические роли>, <эмоциональные окраски>;*

**<леммы фразеологизма> ::=** *<лемма слова> {<лемма слова>}*

**<значения> ::=** *<значение> {, <значение>}*

**<значение> ::=** *<слово> {<слово>}*

**<флаг изменяемости> ::=** *0| 1*

**<синтаксические роли> ::=** *<синтаксическая роль> {, <синтаксическая роль>}*

**<синтаксическая роль> ::=** *<число>*

**<эмоциональные окраски> ::=** *<эмоциональная окраска> {, < эмоциональная окраска>}*

**<эмоциональная окраска> ::=** *<число>*

***<число> ::= 1| 2| 3| ...***

Некоторые дополнительные примеры преобразования лингвистических словарных статей в электронные:

- «**даром хлеб есть**. *Разг., неодобр*. Быть без дела, занятия, не приносить пользы.» - *«[даром, хлеб, есть] : [«Быть без дела», «Не приносить пользы»], 1, [], [4, 11]»*;
- «**не моргнув глазом**. *Обст. Разг*. Не раздумывая очень долго; не поддаваясь страху.» - *«[не, моргать, глаз] : [«Не раздумывая очень долго», «не поддаваясь страху»], 0, [3], [15]»*;
- «**с царём в голове**. Сообразителен, умен.» - *«[с, царь, в, голова] : [«Сообразителен», «Умен»], 0, [], []»*;
- «**яблоко раздора**. *Книжн*. Повод, причина конфликта.» - *«[яблоко, раздор] : [«Причина конфликта», «Повод конфликта»], 1, [], [3]»*;

**Описание программного модуля.** Для работы со словарем фразеологизмов разработан на языке программирования Python 3.3.5 [2] программный модуль, в котором и происходит взаимодействие пользователя со словарем. Можно пополнять словарь, просматривать содержимое и очищать словарь.

При выборе пункта «Добавить в словарь фразеологизмы» управление передается функции, в которой происходит открытие нового окна, в котором можно записать сам фразеологизм, его значения, синтаксические роли в предложении и эмоциональные окраски, а также отметить, изменяем фразеологизм или нет. Как уже было отмечено, фразеологизм может не иметь определенных синтаксических ролей и эмоциональных окрасок.

Когда пользователем введены все данные по фразеологическому обороту, и он решает добавить данные в словарь, управление передается другой функции. Ее аргументы – заполненные или пустые поля окна ввода. В этой функции происходит открытие и изменение словаря (если введенный фразеологизм уже находится в словаре).

Значения, синтаксические роли и эмоциональные окраски добавляются как есть, они разделены символом переноса строки. Сам же фразеологизм дополнительно обрабатывается морфологическим анализатором, для того, чтобы получить леммы слов, входящих в его состав, и ключевое слово фразеологизма. Ключевые слова необходимы для быстрого автоматического поиска фразеологизмов и для уникальности фразеологизмов. Это первые глаголы или существительные, встреченные в фразеологизме при разборе.

Структура XML-файла словаря получается следующей:

```
<dictionary>
    <keys>
        <key word="рыба">
            <phrase full="ни рыба ни мясо">
                <key_position>1</key_position>
                <is_changeable>1</is_changeable>
                <tokens>
                    <lemma>ни</lemma>
```

```
                    <lemma>рыба</lemma>
                    <lemma>ни</lemma>
                    <lemma>мясо</lemma>
                </tokens>
                <meanings>
                    <meaning>Заурядная личность</meaning>
                    <meaning>Посредственная
                личность</meaning>
                </meanings>
                <roles>
                    <role>2</role>
                </roles>
                <emotions>
                    <emotion>4</emotion>
                    <emotion>11</emotion>
                </emotions>
            </phrase>
        </key>
    </keys>
    <about>this is phrase dictionary made for my diploma</about>
</dictionary>
```

Словарь состоит из нескольких ключей, для каждого ключа может быть несколько соответствующих фразеологизмов, внутри каждого фразеологизма есть сущности: позиция ключа в фразе, флаг изменяемости, леммы слов, значения, синтаксические роли и эмоциональные окраски.

**Заключение:** Некоторые правила пунктуации русского языка можно проверить на этапе синтаксического анализа с помощью словарей. Разработана структура и предложена формализация словаря фразеологических оборотов. Полученный электронный словарь фразеологических оборотов можно использовать при создании различных систем автоматического синтаксического и семантического анализа, анализа тональности и генерации текстов.

**Литература:**

1. Ю. А. Ларионова. Фразеологический словарь современного русского языка. – М.: Аделант, 2014. – 512 с.

2. Лутц М. Программирование на Python, том I, 4-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 992 с.

***Abstract:*** *Punctuation is important because it helps people to understand the meaning of writing, get writer's point clear. Some rules require a dictionary of idioms to be checked. There is a linguistic dictionary of Russian idioms, but it is inapplicable in natural language processing, so electronic version of this dictionary must be designed. This article formalizes the existing linguistic dictionary entries with the identification of the necessary attributes for automatic processing in texts. For the convenience of using the electronic dictionary a software module in python 3.x was made. The dictionary was created as XML-file. This dictionary can be used in various systems of*

*automatic syntactic and semantic analysis, sentiment analysis and text generation.*

**Key words:** natural language processing, electronic dictionary, automatic checking of punctuation, formalized electronic idioms dictionary.

***References:***

1. J. A. Larionova. Dictionary of idioms in modern Russian language. – M.: Adelant, 2014.– 512p.

2. Lutz M. Programming Python, vol. I, 4-th edition. – Translation from English – SPb.: Simvol-Plus, 2011. – 992 p.

Статья отправлена: 28.05.2018 г.