Multilingual dictionary of Slavic somatic phrasemes

Masha Bekjкovikj
Faculty of Computer Science and Engineering
University St Cyril and Methodius
Skopje, Macedonia
bekjkovikj.masha@students.finki.ukim.mk

Elena Markova
Faculty of Computer Science and Engineering
University St Cyril and Methodius
Skopje, Macedonia
markova.elena@students.finki.ukim.mk

Katerina Zdravkova
Faculty of Computer Science and Engineering
University St Cyril and Methodius
Skopje, Macedonia
katerina.zdravkova@finki.ukim.mk

***Abstract*—Phrasemes are multi-word expressions and utterances with a distinctive syntax and semantics. Due to the ethnological diversity, they are language specific and in many occasions, the exact translation of a phraseme existing in one source language doesn't occur in other target languages, no matter their linguistic and cultural similarity. The presentation of source phrasemes, their meaning and examples in a form of an interactive and searchable multilingual dictionary is a good starting point for further linguistic research. This paper presents the creation of such a dictionary, its basic functionalities, the process of its development, and finally illustrates its performance according to assigned user roles. The paper ends with the further stages of its development.**

Keywords—somatic phrasemes, morphological search, stemming, search

I. Introduction

Phrasemes are linguistic signs consisting of several lexemes. They are represented as ordered triples: signified (the meaning of the phraseme), signifier (the phonetic form), and syntactic (the set of data referring to its co-occurrence with other signs) [1]. Similarly to multi-word expressions and multi-morphemic utterances, phrasemes are groups of lexemes with a meaning distinctive from the meaning of individual lexemes they are obtained from [2]. Therefore, apart from their purely syntactic description, phrasemes have a significant semantic distinction, making them a very fruitful field for linguistic research. The term phraseme usually encompasses the pragmatic and the semantic phrasemes, which are classified into pragmatemes, idioms, collocations and quasi-idioms [1]. Whenever among its constituent words, some body parts of internal organs are included, such linguistic signs become somatic [3]. They can be: somonymic (denoting parts of the human body), osteonymic (skeletal system), angionymic (circulatory system), scplanchonymic (internal organs), sensonymic (senses), and general body lexis (the body) [3].

Due to the ethnological diversity, phrasemes are language specific and in many occasions, the exact translation of a phraseme existing in one source language doesn't occur in other target languages, no matter their linguistic and cultural similarity. They have been extensively explored at bilingual [4], and multilingual level (https://syd.korpus.cz/). The best tool for aligning parallel corpora is undoubtedly GATES [5].

In spite of the variety of aligning tools and environments intended to present the multilingual corpora, none has tackled the aligning of phrasemes, mainly because the matching is affected by the cultural differences of the paralleled languages. In order to produce a correct idiomatic translation, it is necessary to delve deeper into culture, customs and beliefs of the region this expressions originate from. The linguistic and comparative analysis of the phrasemes is the main goal of a recent University project submitted by the Chair of Slavistics within the Faculty of Philology in Skopje. The project intends to contrast and present the behavior of the Slavic somatic phrasemes.

One of its crucial parts of the project is the visual presentation of parallel aligned corpora in a form of an interactive multilingual dictionary. It is based on the collected Macedonian source phrasemes, their meaning, which are enriched with the typical examples extracted from the literature. Each Macedonian phraseme is then associated with the corresponding target translations into Russian, Polish and Czech.

In this paper we present the development and the current stage of the multilingual phraseme dictionary. The structure of the paper is the following: in the Section II, use case analysis of the system is done, presenting the roles of the users and the major interactions they are authorized to perform. Section III presents the creation the dictionary. Search modules, which were one of our major obstacles are described in Section IV. Section V is a small user manual illustrated with the distinctive functionalities. Finally, the conclusions and further development of the system is presented.

II. Use case analysis

The multilingual dictionary is intended to be used by the linguist specialists, as well by users who are interested to research the phrasemes, their meaning and examples. Since most of them have only the basic computing skills, we decided to make the whole system Web based, exceptionally intuitive and user-friendly. In order to be able to create such system, we decided to build it using the open source content management system Drupal [6], which offered us the opportunity to create a powerful, modular, effective and reliable application. The application is hosted on a virtual machine and is running on Acquia Dev Desktop [7].

The system itself comprises four types of users:

- System administrators, authorized to: create the whole content in all the languages existing in the parallel corpora; approve the modifications initiated by language editors and registered users; enable the extension of the multilingual dictionary with new languages; assign the roles of the authorized users, which belong to next two types. This task will be maintained by the creators of the system.
- Language editors, which have an access to all the contents they are responsible to edit, including the right to add, modify and remove the existing phrasemes, their semantics, as well as their representative examples. This task is assigned to language specialists. In order to eliminate the risk of unintentional damage of the contents, modifications by the registered users will be temporarily stored, and become effective only after the approval by a language editor.
- Registered users capable of profound search of the phrasemes and their meaning. They can access the contents, search it using the keywords existing in the source language, extract the target phrasemes and download the final results. Registered users are also able to add or modify existing contents, but these actions are to be approved by the system administrator or language editors.
- Ordinary users have an authorization to search through the whole contents using the keywords existing in the source language and see the target translations.

The following two sections explain the implementation of the system, with particular emphasis to the realization of the search option, which was our main challenge.

III. Implementation of the system

We used a virtual machine with a 2.27 GHz QEMU Virtual CPU version processor, with installed memory (RAM) of 8 GB and 64-bit Windows Operating System. On this virtual machine we installed Acquia Dev Desktop to run our Drupal application on. Acquia Dev Desktop includes all the necessary elements for the needs of our application: Dev Desktop App - for managing the Drupal application; Apache web server [8]; Percona MySQL database server [9]; PHP - programming language that powers Drupal [10]; and finally, phpMyAdmin - for MySQL management and querying [11].

All these elements were properly installed and configured during the installation of Acquia Dev Desktop. On top of this combination of developing environments, we installed the Drupal application. Whilst setting and configuring the Drupal installation, all the parameters were set for the database name, users, passwords, the site name etc. For this application, we used Drupal version 7.42 although there is a newer 8.0.5 version. The reason we decided to use the older version is because all the modules that we wanted and needed to use are not yet implemented or they exist only in the beta version for Drupal 8.

We decided to start developing this application by first concentrating on the design. We chose a theme which is free and available for downloading, as a base for the design, that we thought was appropriate for this type of application, but modified it according to our needs.

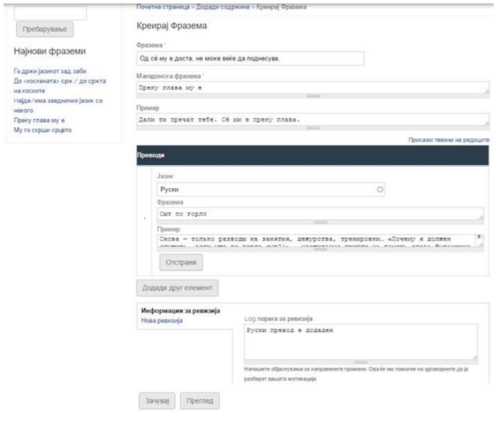
Knowing that we will use multiple languages for the translations of the phrasemes, we defined a taxonomy – vocabulary called *Language* which contains the different languages as terms. This vocabulary can be edited and enriched with new languages at any time without altering any other part of the application.

To create the multilingual phraseme lists. We needed to define a custom content type with custom defined properties. We named this content type *Phraseme* and it is consisted of the following properties: a title which represents the meaning of the phraseme, the original Macedonian phraseme with an appropriate example of usage and a field collection, implemented using the Field Collection module, of all translations. Each translation contains the language of the corresponding phraseme (represented as a vocabulary term which we previously discussed), the corresponding phraseme in that language and an example of its usage in the given language.

To functionally test the accuracy and representation of the content type, we added a few entries of type *Phraseme*. Then we created a tabular view of all *Phraseme* entries. In this view we implemented sorting by column as an option of the table. Also, the five newest phrasemes are displayed as a block on every page. This functionality is implemented so that regular users can easily see if there are new phrasemes added. After we were ensured that this content type and the vocabulary are working as expected, we implemented the Search Option, explained in Section IV.

Defining the types of users and their roles and permissions was implemented last due to the fact that we wanted to be assured that every other functionality of the application works correctly. We defined four types of users as explained in Section II. The system administrator is assigned when installing the Drupal application. This type of user has full access to every aspect of the application, its structure, appearance, people, modules, configuration, content and reports. This user has access to all administer settings. The system administrator can add every type of content available or create new content types, while the other users can only modify or add entries of the content type *Phraseme* and the predefined content type Article (in case they needed to add any news or articles on the home page). Example of this action is shown on Figure 1, which is presented on the next page.

Drupal includes a so called Revisioning module, which is responsible for the configuration of workflows to create, moderate and publish content revisions. By using it we enabled every modification and addition of new content to be sent as a revision to the system administrator or language editor and be published after their approval. When making a revision there is an option to add a revision log message. This way it is easier to trace the changes and explain the motivation to the system administrators and language editors for making the change.



Креирај Фразема

Фразема *
Од сè му е доста, не може веќе да поднесува.

Македонска фразема *
Преку глава му е

Пример
Дали ти пречат тебе. Сè ми е преку глава.

Јазик
Руски

Фразема
Сыт по горло

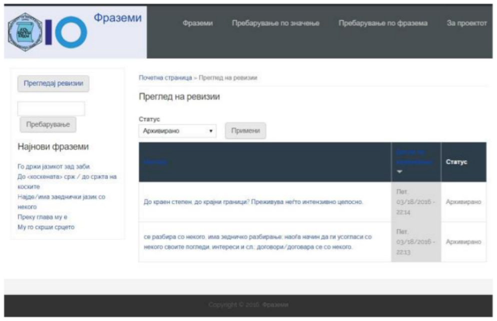
Пример
Снова – только разводы на занятия, дежурства, тренировки… «Почему я должен

Log порака за ревизија
Руски превод е додаден

Напишете објаснување за направените промени. Ова ќе им помогне на одговорните да ја разберат вашата мотивација

Fig. 1. Example of the action: Adding a new phraseme

The system administrator and the language editors have access to a page that displays revision summary for all content. Example of this page is shown on Figure 2. This way there is a better preview of all the revisions – archived, pending and currently published.



Преглед на ревизии

Статус
Архивирано

[illegible]	[illegible]	Статус
До краен степен, до крајни граници? Преживува нешто интензивно целосно.	Пет, 03/18/2016 - 22:14	Архивирано
се разбира со некого, има заедничко разбирање; наоѓа начин да ги усогласи со некого своите погледи, интереси и сл.; договори/договара се со некого.	Пет, 03/18/2016 - 22:13	Архивирано

Fig. 2. Example of the Revision summary page.

Language editors have the permission to extend the corpus with new phrasemes, their explanation and the typical examples explaining the phraseme in Macedonian, together with the translations in all the languages existing in the system. Furthermore, they can add a completely new language and add the corresponding contents. That language becomes available for the phrasemes that were previously entered in the corpus.

To prevent the accidental removal of existing content, the language editor and registered user do not have the option to delete content, rather only to "unpublish" it. The system administrator has access to all content that has been unpublished and has the permission to delete or republish it. Unpublished content is then stored in the database if at any point this data is considered to be restored or published.

We used the Pathauto module to add URL aliases for the application. For example, for phraseme the URL alias is frazema/[node:nid]; for article it is article/[node:title]; for vocabulary terms it is [term:vocabulary]/[term:name] and for users it is users/[user:name].

At this stage of the project, the whole application is based on the dictionary where the Macedonian phrasemes are the core content, and consequently, the major users will be Macedonian native speakers. Therefore, we thought it is more natural and appropriate to have the whole navigation presented in Macedonian.

To enable this navigation, we localized the application by translating the complete Content Management System, its commands, actions, functionalities, buttons etc. using the Translate Interface, in Macedonian so the language editors, registered users and ordinary users would easily navigate on the application.

IV. Search option

The search option, which is the main functionality for all types of users includes the possibility to perform the search using the exact match of the lemmas existing in the source phrasemes, their derivations, but also the parts of the words. This functionalities already exist for the English search. In order to enable it for the Macedonian, we used the electronic lexicon with more than 60K lemmas and their word forms [13].

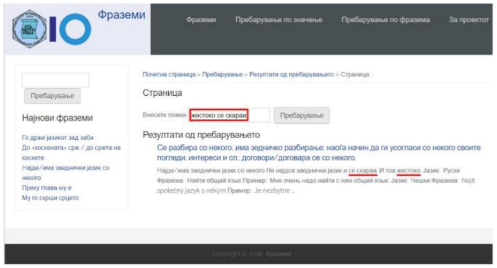
We decided to install and configure the search server on the same virtual machine as the Drupal application because this way the application and the search server communicate locally thus providing better, faster and safer search on the site content. The search server we decided to use is Apache Solr Server 5 (http://lucene.apache.org/solr/). Drupal enables easy connection and communication with this server via modules like Apache Solr Search and Search API.

The Apache Solr Server is a standalone full-text search server powered by the Lucene search library at its core, which is a full-featured text search engine library written entirely in Java. By creating near real-time indexes for the content it is intended to search, Solr is able to achieve fast search responses. This means that Solr does not search in the content directly but in the indexes. Its comprehensive administration interface allows easy access and control over our Solr instance. On the other hand, Apache Server is more commonly used on a Unix-like systems, it was a challenge to configure and run it on a Windows Virtual Machine, also taking in account the fact that on top of an Apache Server configuration we had to configure Solr as well. We had to add Apache Ant in order to compile and run the Solr Server.

Apache Solr Server and Lucene are mainly built and have complete functionality for the English Language. It also has support for 32 languages, including English, but it does not provide support for search content in Macedonian. Due to the fact that both Apache Solr Server and Lucene library are open source, we were able to modify and configure the search to our needs and properly and correctly work for Macedonian language.

In our application, there are three types of searches implemented. They are briefly explained and illustrated below.

The first search is available as a block for every page and uses the Apache Solr Search module. It indexes and searches all the content of the application. This search is really useful if users want to find something fast or are not sure what exacatly are they searching for. Example of this search is shown on Figure 3. The other two searches are implemented with the Search API module because this module enabled us to create custom indexes which are used in custom views.



Почетна страница » Пребарување » Резултати од пребарувањето » Страница

Страница

Внесете поими: жестоко се скараа

Резултати од пребарувањето

Се разбира со некого, има зедничко разбирање: наоѓа начин да ги усогласи со некого своите погледи, интереси и сл.; договори/договара се со некого.

Најде/има заеднички јазик со некого Не најдоа заеднички јазик и се скараа. И тоа жестоко. Јазик: Руски Фразема: Найти общий язык Пример: Мне очень надо найти с ним общий язык. Јазик: Чешки Фразема: Najít společný jazyk s někym Пример: Je nezbytné ...

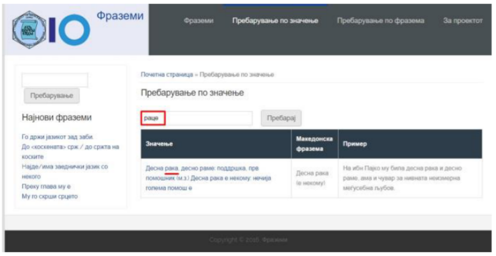
Најнови фраземи

Го држи јазикот зад заби.
До <коскената> срж / до сржта на коскиte
Најде/има заеднички јазик со некого
Преку глава му е
Му го скрши срцето

Copyright © 2016. Фраземи

Fig. 3. Example of the search for all content.

The second search is for searching by the meaning of phrasemes therefore it indexes only the field meaning of the content type Phraseme. This will help users to explicitly search the meanings of phrasemes and will give them a clear view of the original phrasemes, their meaning and an example of usage. Example of this search is shown on Figure 4.



Почетна страница » Пребарување по значење

Пребарување по значење

раце

Значење	Македонска фразема	Пример
Десна рака, десно раме: поддршка, прв помошник (м.з.) Десна рака е некому: нечија голема помош е	Десна рака (е некому)	На ибн Пајко му била десна рака и десно раме, ама и чувар за нивната неизмерна меѓусебна љубов.

Најнови фраземи

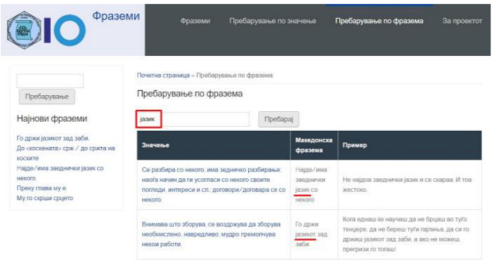
Го држи јазикот зад заби.
До <коскената> срж / до сржта на коскиte
Најде/има заеднички јазик со некого
Преку глава му е
Му го скрши срцето

Copyright © 2016. Фраземи

Fig. 4. Example of search by meaning.

The third search is intended for searching by the original Macedonian phraseme. Therefore it indexes only the field Macedonian phraseme of the content type *Phraseme*. This way users can search for concrete Macedonian phrasemes and similarly as the second search, it provides a tabular view of the original phrasemes, their meaning and an example of usage. Example of this search is shown on Figure 5.

The functionalities of the ordinary users and the registered ones are restricted to preview and search options only. They can go through all the existing contents, as presented on the Figures 3, 4 and 5. While the registered users have the permission to download and print the contents, these activities are disabled to ordinary users.



Почетна страница » Пребарување по фразема

Пребарување по фразема

јазик

Значење	Македонска фразема	Пример
Се разбира со некого, има зедничко разбирање: наоѓа начин да ги усогласи со некого своите погледи, интереси и сл.; договори/договара се со некого.	Најде/има заеднички јазик со некого	Не најдоа заеднички јазик и се скараа. И тоа жестоко.
Внимава што зборува, се воздржува да зборува необмислено, навредливо; мудро премолчува некои работи.	Го држи јазикот зад заби.	Кога еднаш ќе научиш да не брцаш во туѓо тенџере, да не береш туѓи гајлиња, да си го држиш јазикот зад заби, а ако не можеш, прегризи го тогаш!

Најнови фраземи

Го држи јазикот зад заби.
До <коскената> срж / до сржта на коскиte
Најде/има заеднички јазик со некого
Преку глава му е
Му го скрши срцето

Fig 5. Example of search by Macedonian phrasemes.

V. Multilingual dictionary at work

The application is already stable, and it has been approved by the colleagues from the Faculty of Philology. After they test it more exhaustively, it will be available for the wider audience.

As visible from the figures presented in the previous two sections, the application has a very intuitive interface, so the users who are not very familiar with the technology can use it without any assistance.

The actions of the registered and ordinary users have been explained and illustrated in details so far. They can enjoy the wealth of the phrasemes, idioms, and phrases in many languages.

The language editors are currently exhaustively exploring the application. They have provided us with a very small pilot corpus intended to be used for the creation purposes. They have already started with the manual enlargement of the corpus with many new entries, and they also try to extend the multilingual aspect, wherever possible. Although their feedback is very favorable, we are very excited to see whether all of them will be able to use it.

We are steadily performing the system administrator's tasks by carefully observing their work and approving the modifications they produce. So far, the use of the application is very smooth.

VI. Conclusions and further extensions

Multilingual dictionary of Slavic somatic phrasemes is the first project that presents the collected Macedonian idiomatic expressions and compares them with the translation equivalents existing in Russian, Polish and Czech. It was tested and approved by the colleagues from the Faculty of Philology, who found the system intuitive and easy to use.

The system was created with an open source content management system, thus the contributed files, as well as our derived work are licensed under the GNU General Public License, version 2 [14] or later and copyrighted under the Creative Commons Attribution-ShareAlike 2.0 Generic [15]. However, the content of the multilingual dictionary is a property of the team from the Faculty of Philology and can't be downloaded and reproduced without their permission.

In the further stage of the project, search will be extended to enable the possibility for suggestion of the lemmas and their word forms. Currently, the whole navigation in the system, as well as the search option are restricted to Macedonian language only.

We have already enabled the multilingual navigation in all other languages. The enlargement of the search option will depend on the profound linguistic support of language editors.

You are all welcome to explore the application too, and to send us your professional opinion. Thank you in advance.

REFERENCES

[1] I. Mel'čuk. "Collocations and lexical functions." In A. P. Cowie (ed.), "Phraseology: Theory, analysis, and applications. Clarendon Press, 1998, pp. 23-54.

[2] B. Salehi, N. Mathur, P. Cook and T. Baldwin. "The impact of multiword expression compositionality on machine translation evaluation." In Proceedings of NAACL-HLT, 2015, pp. 54-59.

[3] M. Němcová, "Comparative analysis of English and French body Idioms." PhD diss., Masarykova univerzita, Pedagogická fakulta, 2013.

[4] C. Bannard, and C. Callison-Burch. "Paraphrasing with bilingual parallel corpora." In Proceedings of the 43rd Annual meeting on association for computational linguistics, association for computational linguistics, 2005, pp. 597-604.

[5] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan, A framework and graphical development environment for robust NLP tools and applications, 2002, pp. 168-175.

[6] J. Arumugam, R. Parthasarathy and L. Yacub, "Drupal as a content management system in libraries: A study". In Indian journal of science, 21(72), 2015, pp.184-188.

[7] Acquia Dev Desktop, available from https://www.acquia.com/products-services/dev-desktop

[8] Apache web server, , available from https://httpd.apache.org/

[9] Percona MySQL database server, available from https://www.percona.com/software/mysql-database/percona-server

[10] PHP - programming language that powers Drupal, available from http://php.net/

[11] phpMyAdmin - for MySQL management and querying, available from https://www.phpmyadmin.net/

[12] Metro Zymphonies Theme, available from https://www.drupal.org/project/metro_ zymphonies _theme

[13] A. Petrovski, "Morphological computer lexicon – contribution to Macedonian language resources" (in Macedonian), PhD thesis, 2008, University St. Cyril and Methodius, Faculty of Natural Sciences and Mathematics

[14] GNU General Public Licence, available from http://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html

[15] Creative Commons, available from http://creativecommons.org/licenses/by-sa/2