Lexical Database of the Experimental Bulgarian–Polish Online Dictionary *

Ludmila Dimitrova[1], Rumyana Panova[2], Ralitsa Dutsova[2]

[1] Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences, Sofia
[2] Veliko Tarnovo University & IMI-BAS Master Program,
Sofia, Bulgaria

Abstract. In this paper we describe briefly the experimental ongoing version of the Bulgarian–Polish online dictionary. We focus our attention to the lexical database of the dictionary. The starting point for the formal model of lexical database of the dictionary is the CONCEDE model for dictionary encoding. Thus the first Bulgarian-Polish online dictionary will be compatible with other TEI-conformant resources. Some examples from lexical database are presented.

1 Introduction

The base of the first Bulgarian-Polish experimental online dictionary is the ongoing version of the Bulgarian-Polish electronic dictionary [1], [2]. The procedure for selecting the headwords is very simple: we take the headwords from the electronic dictionary. The Bulgarian-Polish electronic dictionary is currently developed in WORD-format in the framework of the joint research project "Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary" between IMI-BAS and ISS-PAS under the supervision of L. Dimitrova and V. Koseska. The current version consists of approximately 20 thousand dictionary entries.

2 Formal model for the Bulgarian-Polish online dictionary encoding

The starting point for the formal model of lexical database (LDB) of the dictionary is the CONCEDE model for dictionary encoding that respect the guidelines of the Text Encoding Initiative Dictionary Working Group (TEI-DWG) [6]. The LDB of the project CONCEDE [4] has standardised and well-understood structure and semantics, and so the first Bulgarian-Polish online dictionary will be compatible with other TEI-conformant resources. With the support of the European Commission the CONCEDE (*Consortium for Central European Dictionary Encoding)* prepared lexical databases for the six Central and East European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene [5]. The first LDB for Bulgarian, more than 2700 lexical entries from the Bulgarian Explanatory Dictionary, based on encoding standards established by the TEI was developed in CONCEDE project [3].

3 Lexical Database

We start to develop the structured LDB taking the recent version of the ongoing Bulgarian-Polish electronic dictionary. This LDB is an entry point to the relational database (RDB) of the Bulgarian-Polish online dictionary. Whenever possible the LDB will generate a new structure of entries for the Polish-Bulgarian online dictionary.

* The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX

The ***structural tags***, used in the LDB of the Bulgarian-Polish online dictionary, are three: **entry, struc, alt.**

alt: alternation, though generally for use in quite different contexts
entry: dictionary entry
struc: indicates separate independent part in the dictionary entry.

The set of ***content tags*** includes the elements:

case: contains grammatical case information given by a dictionary for a given form
conjugation: *a new tag* is added to represent the conjugation of verbs; its structure allows the sub tag **type** for the possible types of conjugations of Bulgarian verbs
def: directly contains the text of the definition
domain: domain
eg: a structure, contains an example, as given in a dictionary, and allows the tags **source** and **q**
etym: a structure, contains etymological information and allows the tags **lang** and **m**, as given in a dictionary
gen: identifies the morphological gender of a lexical item, as given in the dictionary
geo: geographic area
gram: contains grammatical information relating to a word other than gender, number, case, person, tense, mood, itype, as these all have their own element, for example, perfect aspect and progressive aspect
hw: the headword; used for alphabetization and indexing, access
itype: indicates the inflectional class associated with a lexical item, as given in a dictionary
lang: language; for use in etymologies (in **etym**)
m: indicates a grammatical morpheme in the context of etymology
mood: contains information about the grammatical mood of verbs, as given in a dictionary
number: indicates grammatical number associated with a form, as given in a dictionary
orth: gives the orthographic form of a dictionary headword
person: indicates grammatical person associated with a form, as given in a dictionary
pos: indicates the part of speech assigned to a dictionary headword (noun, verb, adjective, etc.)
q: contains a quotation or apparent quotation
register: register, for type attribute on **usg** tag
source: bibliographic source for a quotation
subc: contains sub-categorization information (transitive/intransitive, countable/non-count, etc.)
time: temporal, historical era, for example, “archaic”, “old”, etc.
type: *a new* subtag in the frame of **conjugation** tag indicates explicitly one of the three types of conjugation of the Bulgarian verbs
tns: indicates the grammatical tense associated with a given inflected form in a dictionary **trans**: contains translation text and related information, so may contain any of the content tags; the principle is that everything under **trans** relates to the target language
usg: contains usage information in a dictionary entry, other than **time**, **domain**, **register** (as these all have their own element), like “dialect”, “folk”, “colloquialism”, etc.
xr: uses to indicate a cross reference with the pointer.

4 Dictionary entry samples

The following samples represent the dictionary entry in XML format and suggest a structure of this dictionary entry in the database of the dictionary to be presented on the Internet. Let us introduce some notation used in the lexical database. We used “’” to mark the accent of the words. The symbol “|” is used to separate the variable part of the word from the main part. The transitive and intransitive verbs should be represented with the corresponding term in the tag **subc**. We introduce “NILL” value in order to represent empty corresponding values.

1) Headword “**притеснение**” /*embarrassment*/

притесне’**ни|е, -я** *n* ucisk *m,* udręczenie *n*, uciemiężenie *n*, przygnębienie *n*; kłopoty materialne

```
<entry>
    <hw>притесне’ни|е</hw>
   <alt>
       <orth>-я</orth>
       <num>pl</num>
   </alt>
    <gen>n</gen>
    <struc type=”Sense” n=”1”>
         <trans>ucisk</trans>
         <gen>m</gen>
         <alt>
                 <trans>udręczenie</trans>
                 <gen>n</gen>
         </alt>
         <alt>
                 <trans>uciemiężenie</trans>
                 <gen>n</gen>
         </alt>
         <alt>
                 <trans>przygnębienie</trans>
                 <gen>n</gen>
         </alt>
    </struc>
    <eg>
         <q>NILL</q>
         <transl>kłopoty materialne</transl>
    </eg>
</entry>
```

2) Headword “поддавам се” /succumb, give way/

подда’ва|м се, -ш *vi.* poddawać się, ulegać, ustępować; **това не се ~ на описание** tego nie da się opisać; **~ ми се нещо** *pot.* coś idzie mi łatwo

```
<entry>
    <hw>подда’ва|м се</hw>
    <pos>v</pos>
    <gram>i</gram>
    <subc>transitive</subc>
    <conjugation>
         <orth>-ш</orth>
         <type>I</type>
    </conjugation>
    <struc type=”Sense” n=”1”>
         <trans>poddawać się</trans>
         <alt>
                 <trans>ulegać</trans>
         </alt>
```

```
        <alt>
                <trans>ustępować</trans>
        </alt>
    </struc>
    <eg>
        <q>~ това не се ~ на описание</q>
        <transl>tego nie da się opisać</transl>
    </eg>
    <eg>
        <q>~ ми се нещо</q>
        <usg type=”register”>pot</usg>
        <transl>coś idzie mi łatwo</transl>
    </eg>
</entry>
```

3) Headword “**притежателен**” /*possessive*/

притежа’тел|ен, -на, -но *adi. gram.* dzierżawczy; **~ни местоиме’ния** zaimki dzierżawcze

```
<entry>
    <hw>притежа’тел|ен</hw>
    <alt>
        <orth>-на</orth>
        <gen>f</gen>
    </alt>
    <alt>
        <orth>-но</orth>
        <gen>n</gen>
    </alt>
    <pos>adi</pos>
    <usg type=”register”>gram</usg>
    <struc type=”Sense” n=”1”>
        <trans>dzierżawczy</trans>
    </struc>
    <eg>
        <q>~ни местоиме’ния</q>
        <transl>zaimki dzierżawcze</transl>
    </eg>
</entry>
```

4) Headword I “**под**” /*under*/, **II** “**под**” / *floor*/

I **под** *praep.* pod; poniżej; **миньорите работят ~ земята** górnicy pracują pod ziemią; **усмихвам се ~ мустак** uśmiecham się pod wąsem; **държа нещо ~ ключ** trzymam coś pod kluczem; **пет градуса ~ нулата** pięć stopni poniżej zera; **парите са вложени в банката ~ лихва** pieniądze są złożone w banku na procent

II **под, -о’ве** *m* podłoga *f*

```
<entry n=”1”>
    <hw>под</hw>
    <pos>praep</pos>
    <struc type=”Sense” n=”1”>
        <trans>pod</trans>
    </struc>
    <struc type=”Sense” n=”2”>
```

```
            <trans>poniżej</trans>
        </struc>
        <eg>
            <q>миньорите работят ~ земята</q>
            <transl>górnicy pracują pod ziemią</transl>
        </eg>
        <eg>
            <q>усмихвам се ~ мустак</q>
            <transl>uśmiecham się pod wąsem</transl>
        </eg>
        <eg>
            <q>държа нещо ~ ключ</q>
            <transl>trzymam coś pod kluczem</transl>
        </eg>
        <eg>
            <q>пет градуса ~ нулата </q>
            <transl>pięć stopni poniżej zera</transl>
        </eg>
        <eg>
           <q>парите са вложени в банката ~ лихва</q>
           <transl>pieniądze są złożone w banku na procent</transl>
        </eg>
</entry>

<entry n=”2”>
        <hw>под</hw>
        <alt>
            <orth>-о’ве</orth>
            <num>pl</num>
        </alt>
        <gen>m</gen>
        <struc type=”Sense” n=”1”>
            <trans>podłoga</trans>
            <gen>f</gen>
        </struc>
</entry>
```

5) Headword “**поддам се**” /succumb, give way/

подд|а’м се, -а’деш *vp.* **v. подда’вам се**

```
 <entry>
        <hw>подд|а’м се</hw>
            <pos>v</pos>
            <gram>p</gram>
            <subc>transitive</subc>
            <conjugation>
                    <orth>-а’деш</orth>
                    <type>I</type>
            </conjugation>
        <xr>подда’вам се</xr>
</entry>
```

5 Relational Database

The model of a relational database is experimentally based on a limited number of studied lexical entries. In the design of the relational database we have provided also the opportunity for translation from Polish to Bulgarian language. That translation will be made only from the main meanings of the Bulgarian headwords. No derivations, phrases or examples will be used for translating from Polish to Bulgarian language.

The relational database is presented on *Figure 1*.

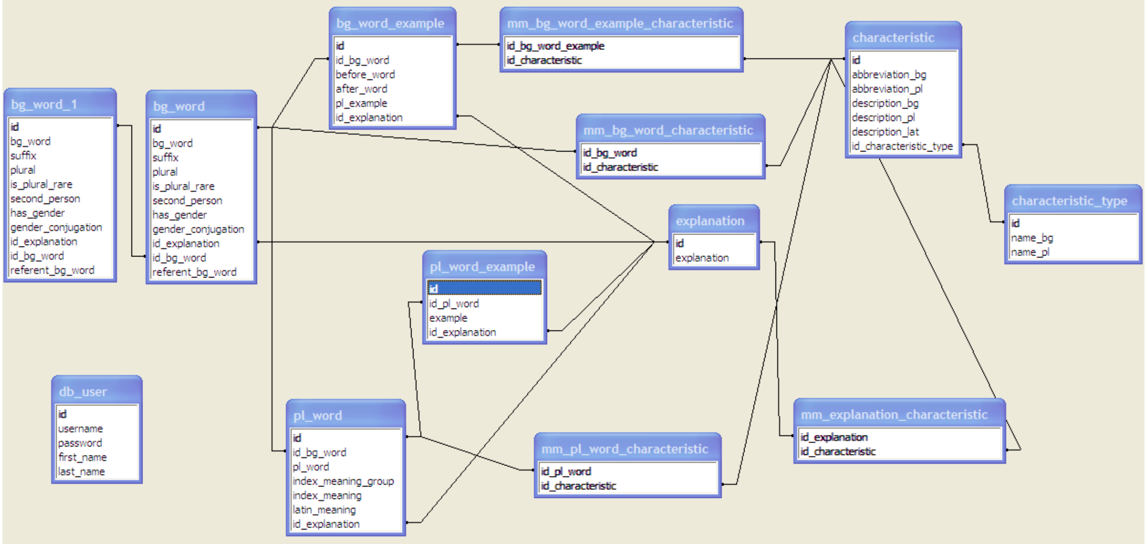


Figure 1: Relational database upon the lexical database of the Bulgarian-Polish-Bulgarian Dictionary

Detailed information on the base units follows.

Table: bg_word

Description: Bulgarian headwords

Field	**Type**	**Null**	**Default**	**Comments**
id	int(11)	No		Id
homonym_index	int(1)	Yes	*NULL*	Index of the homonym (if null, no homonym exists)
bg_word	varchar(100)	No		Bulgarian headword
suffix	varchar(20)	Yes	*NULL*	Suffix
plural	varchar(20)	Yes	*NULL*	Plural form for a noun
is_plural_rare	int(1)	Yes	*NULL*	Frequency of usage of the plural form for a noun (null – normal, 0 - often, 1 – rare)
conjugation	varchar(20)	Yes	*NULL*	Conjugation form for a verb (2 p., present)
conjugation_type	int(1)	Yes	*NULL*	Type of conjugation for a verb (1, 2 or 3)

has_gender	int(1)	Yes	*NULL*	Whether a noun has feminine and neuter gender
gender_feminine	varchar(20)	Yes	*NULL*	Feminine gender form for an adjective
gender_neuter	varchar(20)	Yes	*NULL*	Neuter gender form for an adjective
id_explanation	int(11)	Yes	*NULL*	Foreign key to “explanation”
id_bg_word	int(11)	Yes	*NULL*	Id of the referent Bulgarian word
referent_bg_word	varchar(255)	Yes	*NULL*	Referent Bulgarian word

Table: bg_word_example

Description: Derivations, phrases or examples of the Bulgarian headwords and their translation in polish

Field	**Type**	**Null**	**Default**	**Comments**
id	int(11)	No		Id
id_bg_word	int(11)	No		Foreign key to “bg_word”
before_word	varchar(100)	Yes	*NULL*	Text before the headword
after_word	varchar(100)	Yes	*NULL*	Text after the headword
type	int(1)	No		Type of the usage (1 - Derivation; 2 - Phrase; 3 - Example)
pl_translation	varchar(255)	Yes	*NULL*	Polish translation
id_explanation	int(11)	Yes	*NULL*	Foreign key to “explanation”

Table: pl_word

Description: Polish headwords

Field	**Type**	**Null**	**Default**	**Comments**
id	int(11)	No		Id
id_bg_word	int(11)	No		Foreign key to “bg_word”
pl_word	varchar(100)	Yes	*NULL*	Polish headword
sense_index	int(2)	No		Index of the sense
alternative_sense_index	int(2)	No		Index of the alternative sense
latin_translation	varchar(255)	Yes	*NULL*	Latin translation of the word
id_explanation	int(11)	Yes	*NULL*	Foreign key to “explanation”

Table: pl_word_example

Description: Examples of the polish headwords

Field	**Type**	**Null**	**Default**	**Comments**
id	int(11)	No		Id
id_pl_word	int(11)	No		Foreign key to “pl_word”
example	varchar(255)	No		Example in Polish
id_explanation	int(11)	Yes	*NULL*	Foreign key to “explanation”

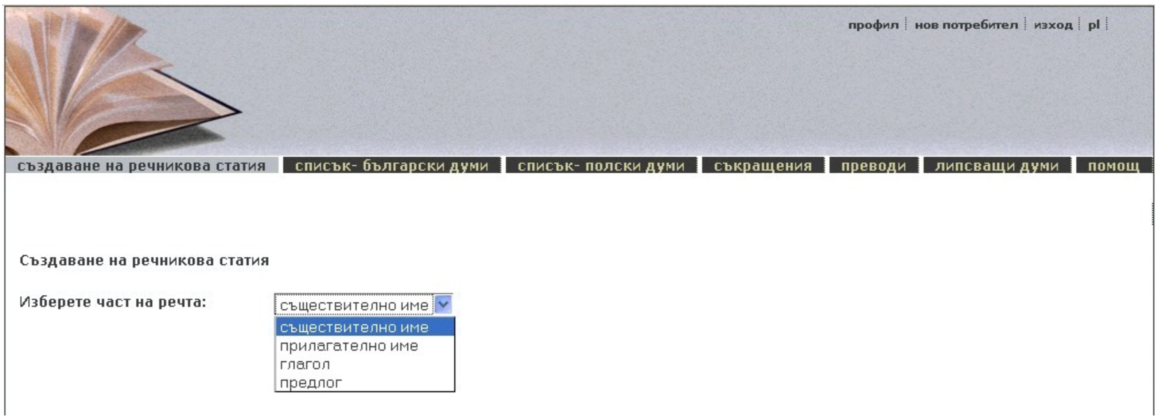
Further improvements will be made when we examine more lexical entries.

6 Web-based Application

The web-based application consists of **administrator and end-user modules**. The administrator module is used to fill in the database and to offer user- friendly interface to the administrators. The idea is that both end-user and administrative parts of the web-based application be bilingual. The following web-based application is experimental, and the structure of the text fields is not permanently determined yet. Changes are possible during the implementation process.

The technologies used for the implementation of the web-based application are Apache, MySQL, PHP and JavaScript. We use free technologies originally designed for developing dynamic web pages with a lot of functionalities. With the help of HTML and CSS we created the designs of both administrative and end user modules. The **administrator module** is intended for the person updating the dictionary. It offers a user-friendly interface for adding, editing, deleting and searching words. The access to the administrative module will be possible only for authorized users. There are possibilities to create more than one user with different passwords and usernames. After the user’s password and username have been verified, the user is redirected to the administrative module where there are **several sections** - **section** for entering a new word, **sections** for searching Bulgarian or Polish words, **section** where the user can enter new abbreviations, **section** for setting translations of the user alerts and messages so the user can change the both Polish and Bulgarian translations, **section where** end-users report the missing words. The Help section serves both the administrative and the end users.

Section for entering a new word: from the beginning the user must choose from a combo box what he wants to enter - noun, verb, adjective or any other part of speech (pronoun, conjunction, adverb). Than with the help of AJAX only the corresponding text fields are loaded.

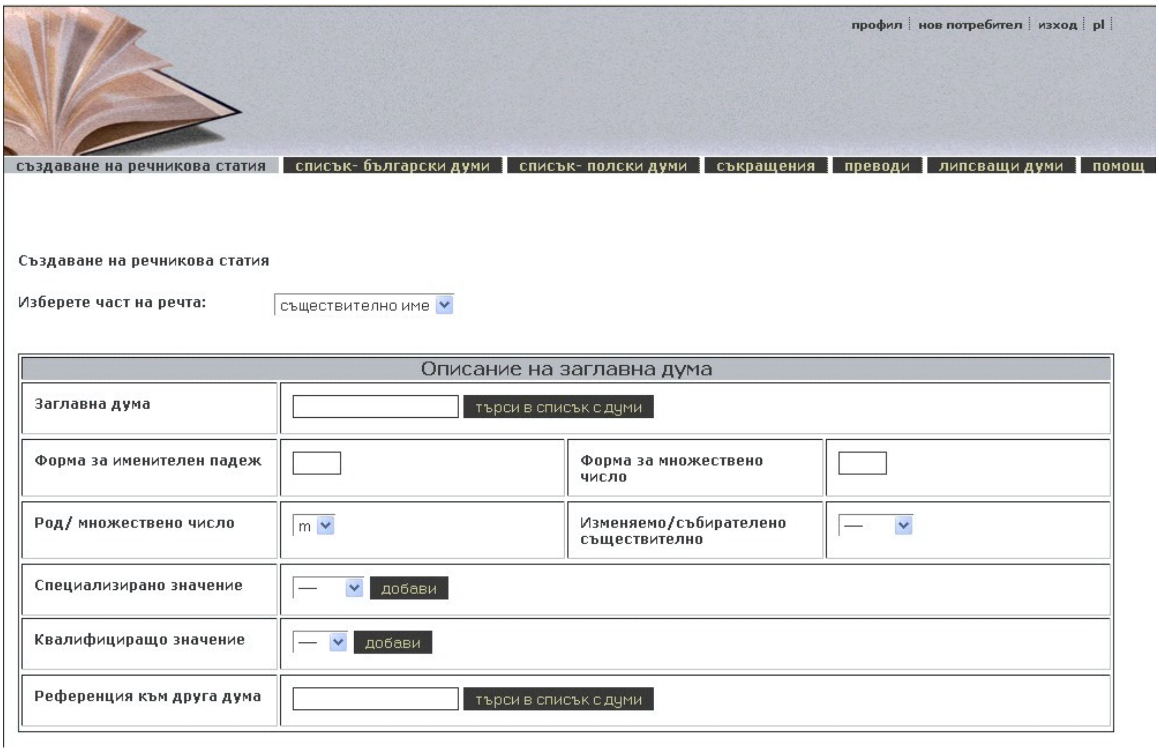


Създаване на речникова статия

Изберете част на речта: съществително име

Figure 2: Administrative panel - choosing the type of the word which will be added

When the user wants to add a **new noun** the fields which are necessary for describing nouns are displayed - field for the headword, combo box for choosing the gender of the noun, etc... With the help of AJAX the user has the opportunity to add as many as needed qualified abbreviations like (archaic, dialect, colloquial) or specialized abbreviations like (botanical, chemistry, anatomy, astronomy).



Създаване на речникова статия

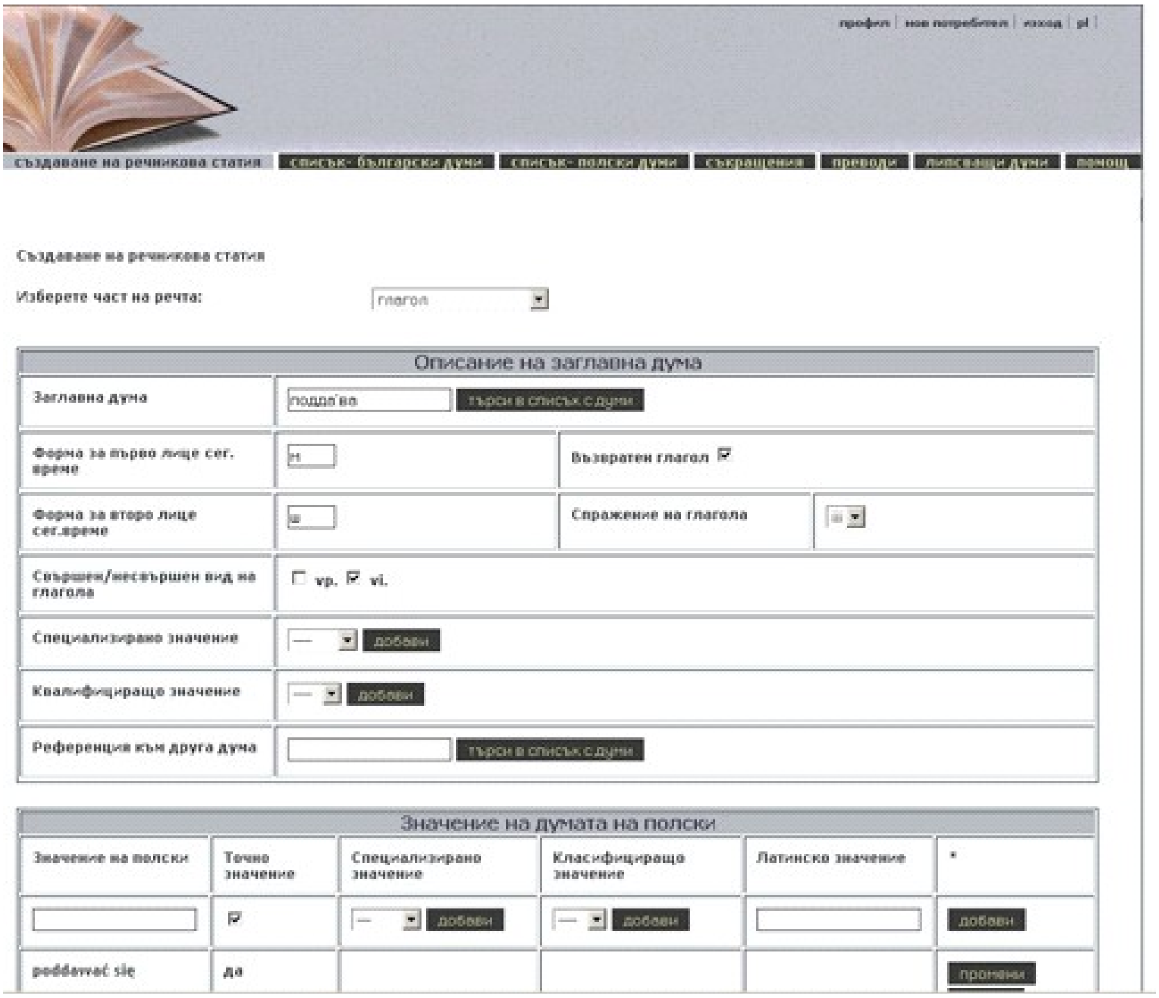
Изберете част на речта: съществително име

Описание на заглавна дума

Заглавна дума	търси в списък с думи		
Форма за именителен падеж		Форма за множествено число	
Род/ множествено число	m	Изменяемо/събирателено съществително	—
Специализирано значение	— добави		
Квалифициращо значение	— добави		
Референция към друга дума	търси в списък с думи		

Figure 3: Administrative panel - adding a noun

When the user adds a **new verb** the displayed fields are headword, checkboxes for choosing perfect aspect (vp) or imperfect aspect (vi) of the verb, etc. To display the conjugation of the verb (except showing the conjugation of the verb in 3rd person, singular) we add an extra field where the user can specify the conjugation type. In the help of the administrative module there is an explanation how to determine the conjugation type of any verb.



Създаване на речникова статия

Изберете част на речта: глагол

Описание на заглавна дума

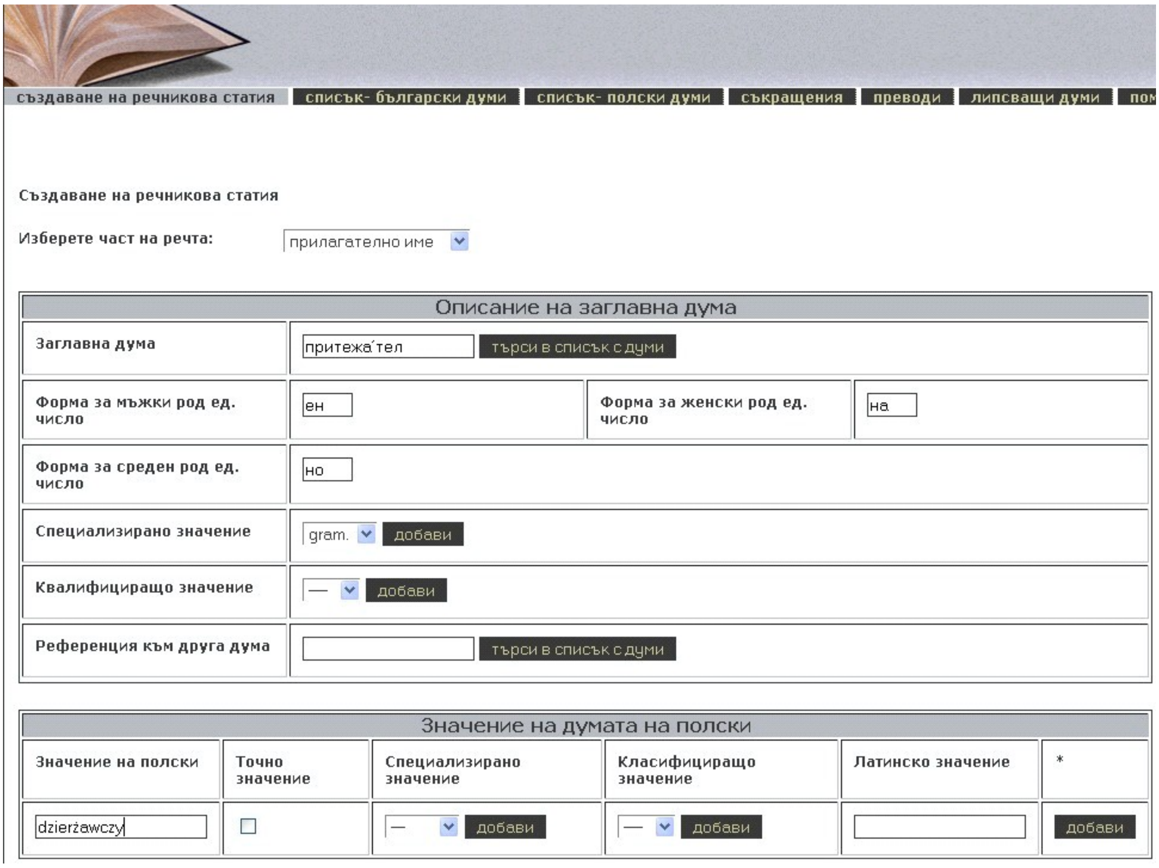
Заглавна дума	подда'ва		
Форма за първо лице сег. време	н	Възвратен глагол ☑	
Форма за второ лице сег.време	ш	Спрежение на глагола	[illegible]
Свършен/несвършен вид на глагола	☐ vp. ☑ vi.		
Специализирано значение	—		
Квалифициращо значение	—		
Референция към друга дума			

Значение на думата на полски

Значение на полски	Точно значение	Специализирано значение	Класифицираща значение	Латинско значение	*
	☑	—	—		
poddawać się	да				

Figure 4: Administrative panel - adding a verb

When adding a **new adjective,** fields specifying the forms for masculine, feminine and neuter are displayed.



Създаване на речникова статия

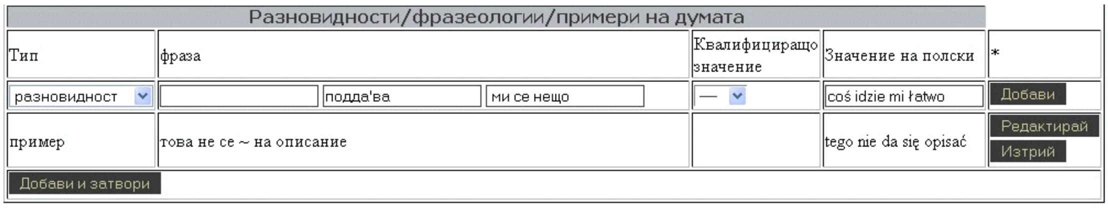
Изберете част на речта: прилагателно име

Описание на заглавна дума			
Заглавна дума	притежа´тел търси в списък с думи		
Форма за мъжки род ед. число	ен	Форма за женски род ед. число	на
Форма за среден род ед. число	но		
Специализирано значение	gram. добави		
Квалифициращо значение	— добави		
Референция към друга дума	търси в списък с думи		

Значение на думата на полски					
Значение на полски	Точно значение	Специализирано значение	Класифициращо значение	Латинско значение	*
dzierżawczy	□	— добави	— добави		добави

Figure 5: Administrative panel - adding an adjective

There is a common part for each part of speech that ensures the possibility to add unspecified number of derivations, phrases and examples for each headword. At the end of each page for entering headword there is a button “Add derivation / phrase / example”. When the user clicks on it a new window is opened in order to add as many as needed **derivations, phrases and examples** for this headword:



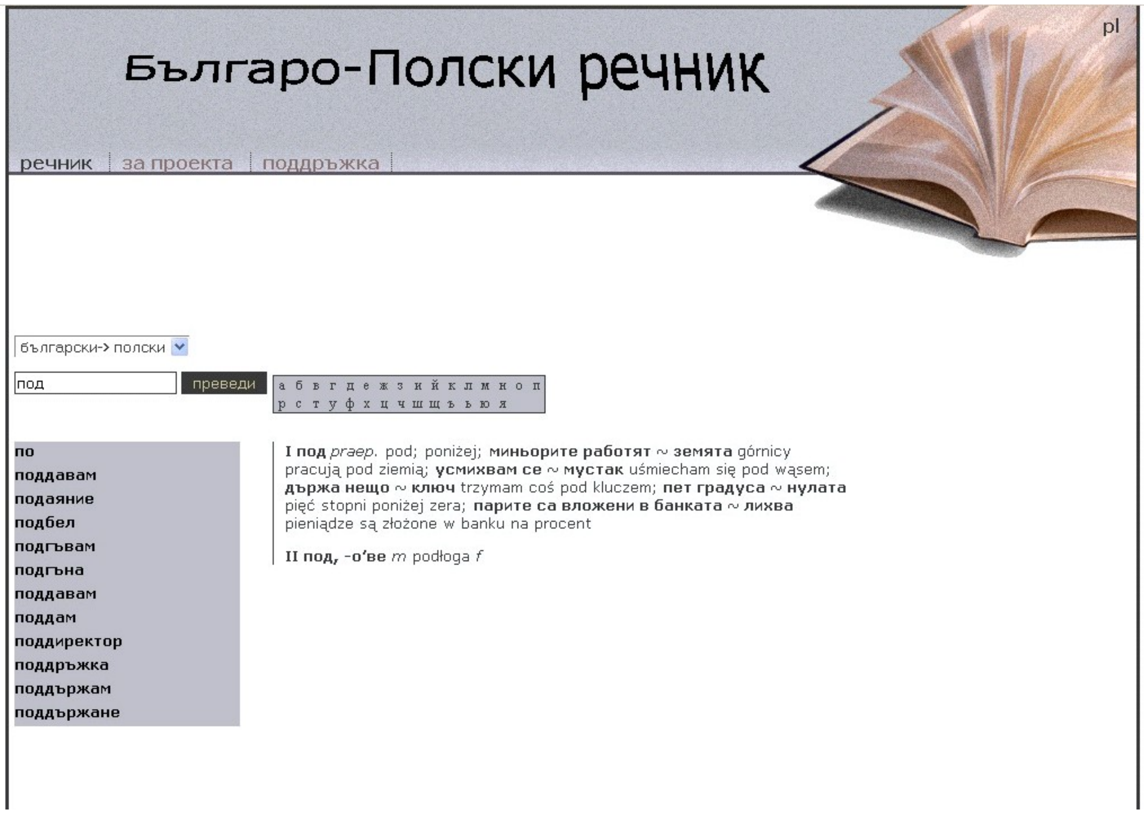
Разновидности/фразеологии/примери на думата				
Тип	фраза	Квалифициращо значение	Значение на полски	*
разновидност	подда'ва ми се нещо	—	coś idzie mi łatwo	Добави
пример	това не се ~ на описание		tego nie da się opisać	Редактирай Изтрий
Добави и затвори				

Figure 6: Administrative panel - adding derivations, phrases and examples for the specified headword

Realization of the homonyms in the web-based application: the meanings of the homonyms are entered in the dictionary as different DB records. In the page for entering the words there is a field where the user must specify a homonym index - a number which shows the order of the meanings.

The web-based end-user application is bilingual as well. In this application there are three sections - section for translating a word, information section and section for reporting a missing word. The user can choose the input language (Bulgarian or Polish) and according to it a virtual Bulgarian or Polish keyboard is displayed. In this way the user can choose special Bulgarian or Polish characters if they are not supported by the different keyboards.

After making a search for a word on the left site of the screen a list of words, starting from the given entry, are displayed. When clicking on any of these words in the list the translation is visualized in the right frame. If we translate from Bulgarian to Polish, the whole information saved in the RDB is displayed. When translating from Polish to Bulgarian, only the Bulgarian headwords are visualized.



български-> полски

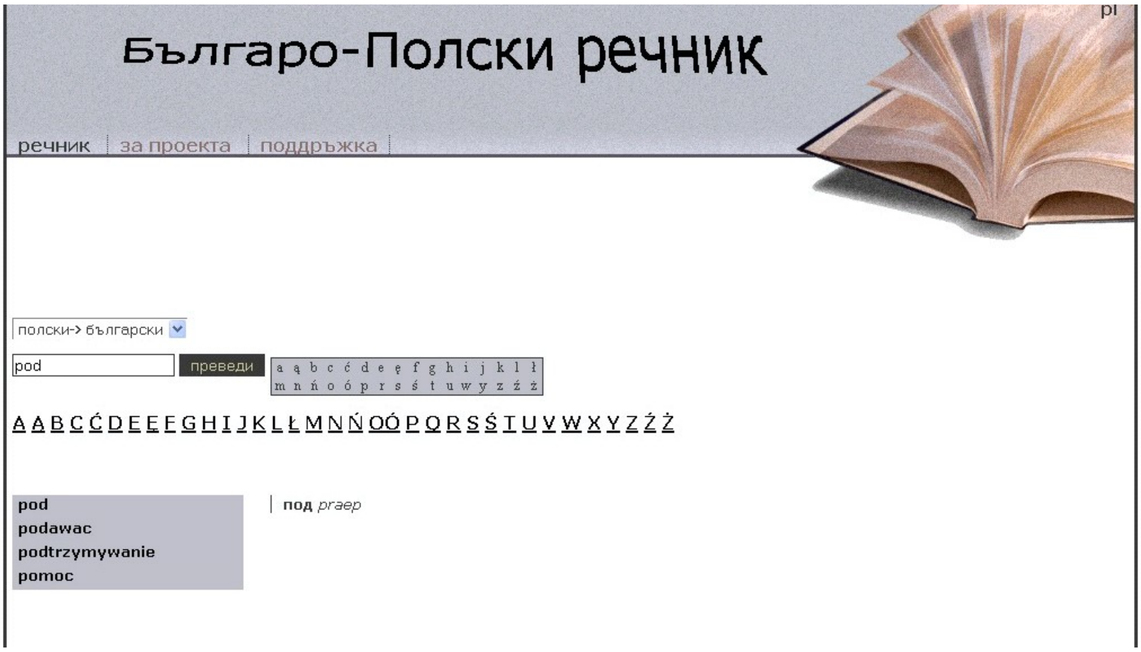
под

по
поддавам
подаяние
подбел
подгъвам
подгъна
поддавам
поддам
поддиректор
поддръжка
поддържам
поддържане

I под *praep.* pod; poniżej; **миньорите работят ~ земята** górnicy pracują pod ziemią; **усмихвам се ~ мустак** uśmiecham się pod wąsem; **държа нещо ~ ключ** trzymam coś pod kluczem; **пет градуса ~ нулата** pięć stopni poniżej zera; **парите са вложени в банката ~ лихва** pieniądze są złożone w banku na procent

II под, -о'ве *m* podłoga *f*

Figure 7: Web page for end users - translation of a Bulgarian word



pl
Българо-Полски речник
речник
за проекта
поддръжка
преведи
a ą b c ć d e ę f g h i j k l ł
m n ń o ó p r s ś t u w y z ź ż
A Ą B C Ć D E Ę F G H I J K L Ł M N Ń OÓ P Q R S Ś T U V W X Y Z Ź Ż

полски-> български

pod

pod
podawac
podtrzymywanie
pomoc

под *praep*

Figure 8: Web page for end users - translation of a Polish word

Both web-based applications have “Help” panels. The end users have the opportunity to report words that are missing in the dictionary into a provided “Contact” form. In this case the administrators will add the reported missing words into the database after.

7 Conclusion

This paper has presented the lexical database of the ongoing version of the first Bulgarian–Polish online dictionary. The formal model of the designed lexical database is CONCEDE model, so the dictionary will be compatible with other TEI-conformant resources.

Due to the limited number of lexical entries taken in consideration, the represented Bulgarian-Polish online dictionary is still at an experimental stage. Further extension of the LDB and RDB will be made.

Bibliography

[1] Dimitrova, L., V. Koseska–Toszewa. (2007). Digital Dictionaries – Problems and Features. Proceedings of the Jubilee International Conference Mathematical and Computational Linguistics. 6 July 2007, Sofia, Bulgaria, pages 25-34.

[2] Dimitrova, L., V. Koseska–Toszewa. (2008). Some Problems in Multilingual Digital Dictionaries. International Journal Études Cognitives. SOW, 8, 237–254.

[3] Dimitrova, L., Pavlov, R., Simov, K. (2002). The Bulgarian Dictionary in Multilingual Data Bases. Cybernetics and Information Technologies, 2(2), 33–42.

[4] Tomaž Erjavec, Roger Evans, Nancy Ide, Adam Kilgarriff. (2000). The Concede Model for Lexical Databases. Proceedings of the Second International Conference on Language Resources and Evaluation, LREC'00. 355-362, ELRA, Paris.

[5] CONCEDE: http://www.itri.brighton.ac.uk/projects/concede/

[6] TEI: http://www.tei-c.org/index.xml