

Chooser User Manual

Table of Contents

1. Introduction.....	1
2. Getting started with Chooser.....	3
2.1. Installation.....	3
2.2. File format.....	3
2.3. Starting Chooser.....	4
3. Overview.....	4
3.1. User interface.....	5
3.1.1. Text view.....	5
3.1.2. List view.....	7
3.1.3. Info view.....	8
3.2. Corpus annotation and editing.....	10
3.2.1. Loading and saving files.....	10
3.2.2. Selecting LUs.....	10
3.2.3. Annotation.....	12
3.2.4. Traversing the corpus.....	13
3.2.5. Corpus searching and editing.....	14
3.2.5.1. Edit operations.....	15
3.2.5.2. Adding/deleting words.....	17
3.2.5.3. Search	21
References	22

1. Introduction

Chooser is an OS-independent multi-functional system for linguistic annotation adaptable to annotation schemata for different language levels. It has been used in the creation of BulPosCor - A POS-annotated corpus of Bulgarian and BulSemCor - a corpus of Bulgarian annotated with wordnet senses http://dcl.bas.bg/en/corpora_en.html. This manual covers the semantic annotation module.

The senses used in the semantic annotation are those encoded in the Bulgarian wordnet. Therefore, Chooser is coupled with the system for wordnet development Hydra

and accesses the senses available in the wordnet database through Hydra's API. Chooser's interface embeds a fully-fledged visualisation of the wordnet synsets. The changes made to the wordnet database are dynamically updated and displayed. Any corrections or additions such as newly created synsets and synonyms are made accessible to Chooser immediately after they are performed. In such a way the semantic annotation takes place simultaneously with the wordnet development.

The basic annotation functionalities implemented in Chooser are:

- (i) fast and easy-to-perform annotation;
- (ii) run-time access to detailed information for the annotation candidates through the associated wordnet senses with all the info pertaining to the respective synsets – synonyms, explanatory definition, semantic relations, notes on usage, grammar, pragmatics, etc.;
- (iii) identification of MWEs with contiguous and non-contiguous constituents;
- (iv) different strategies for corpus traversing;
- (v) operations over the language units in the corpus - edit, insert and delete functions;
- (vi) a flexible search strategy allowing both simple and regular expression search according to wordform or lemma.

All the changes made to the corpus by means of the edit/insert/delete functions are immediately displayed and made accessible in the programme's interface to view and use in the annotation process. The functionalities involving the manipulation of the corpus files minimise considerably the need for prior normalisation.

The system provides multiple-user concurrent access.

Chooser is written in Python and tested under Linux and Windows.

2. Getting started with Chooser

2.1. Installation

For the initial setup of the programme confer the installation manual:

<http://dcl.bas.bg/Tools/Chooser/Chooser-InstallationManual.pdf>

2.2. File format

Pre-processing

The files need to be tokenised and lemmatised. Lemmatisation is essential because a language unit in the corpus is mapped to synsets in wordnet only if the lemma of the LU matches the lemma of a literal in one or more synsets.

File format

The used format is a flat xml. The root element is <text>. The attribute current stores the last position where the file is viewed.

```
<?xml version="1.0" ?><text current="5196">
```

```
...
```

```
</text>
```

The text is encoded as a list of xml elements labeled *word*. The relevant information is stored in separate attributes: wordform (“w”), lemma (“l”), sense (“s”), annotator (“u”), time stamp (“t”), sentence end (“e”). A special attribute is reserved for a parent ID that links the individual tokens of a compound (“p”). An annotated unit contains the following basic information:

<word l="зацова" p="-1529023764" s="1100001720" t="1298483182" w="зацова"/>

Minimum restrictions are imposed on the extension of the specified file format, so that it may permit addition of flat and/or hierarchical annotation schemata without affecting the current one, thus enabling other levels of annotation.

2.3. Starting Chooser

To launch Chooser using a command line, run the following command:

python chooser.py

The examples below show how to run the programme in a Linux environment, provided that it is located in the local directory **/home/boby/chooser**:

(1) from the local directory (chooser) where the executable file **chooser.py** is stored:

boby@tornado:~/chooser\$ python chooser.py

(2) using the full path to the executable file:

boby@tornado:~\$ python /home/boby/chooser/chooser.py

(3) using a relative path:

boby@tornado:~\$ python chooser/chooser.py

3. Overview

The following sections give a description of Chooser's user interface and functionalities.

3.1. User interface

On launching Chooser, the programme's window appears on the screen. It has a tripartite display area consisting of a **Text view** (the main pane), a **List view** (bottom pane) and an **Info view** (right-hand pane). Different types of information are displayed in each field when a file is loaded.



Fig. 1. Chooser on startup

3.1.1. Text view

The **Text view** is the main pane where corpus files are loaded and displayed, as well as

where the corpus is traversed, and the language units are selected for annotation. Many of the operations performed on the language units take place in or are initiated from the **Text view**:

(i) identifying language units (LUs)

A language unit is a word or a group of words that is treated as a single entity, i.e. - assigned a single sense. Usually a token corresponds to a single word LU. However, sometimes the tokens need to be normalised. For instance, numbers may be attached to the following word (e.g. a numbered list in the corpus). For the sake of proper annotation they need to be split, identified, and subsequently annotated, as separate LUs. Chooser supports a run-time edit functionality, which allows such operations to be performed in parallel with the annotation.

(ii) grouping words in multiword expressions (MWE)

Chooser provides a function for MWE identification and grouping, accounting for word order variations of the MWE components and for MWEs with non-adjacent components. For details see the section **Grouping words in multiword expressions**;

(iii) edit operations

Chooser provides a functionality for editing the wordforms and lemmas of the LUs. In such a way typos and wrong lemmatisation do not impede the annotation process. For a description in more detail see **Edit operations**.

(iv) add/remove operations

This feature enables the insertion and deletion of tokens in the corpus files. This is particularly useful in cases of wrong spelling of multiword expression as single words and vice versa. For details see the **Adding/deleting** section.

The status of the LUs with respect to annotation is denoted by means of different colour codes. Several types of units are recognised:

- (a) untraversed units – units in the corpus that have not been traversed in the current session, so no information on their annotation status is displayed;
- (b) annotated units - language units that are assigned a sense;
- (c) non-annotated units - units that have not been assigned a sense;
- (d) compound words/multiword expressions.

3.1.2. List view

The **List view** is a standard list control that displays the definitions of the wordnet synsets suggested as annotation candidates for a LU in the corpus.

The synsets are listed according to decreasing frequency of selection calculated on the basis of the previous choices made by the annotators.

The **List view** and the Info view are synchronised, so that when an item in the list is selected, the corresponding synset is displayed in the Info view.

To browse and select items in the **List view** and view the respective synsets in the Info view, use any of the following methods:

- (i) choose the relevant number (or letter) in front of a given definition in the **List view** using the number or letter keys;
- (ii) browse the **List view** with the **Up/Down Arrow** keys;
- (iii) click on a particular definition in the **List view**.

The **Text view** in Fig. 2 shows the currently selected LU *бизнес* (coloured in red); the

List view displays the definitions of the available wordnet senses - the one highlighted in red corresponds to the wordnet sense that has been assigned to the LU - {бизнес:2; бизнес сектор:1; бизнес сфера:1} (business; business sector). Finally, Hydra's main view for the synset {бизнес:2; бизнес сектор:1; бизнес сфера:1} is displayed in the **Info view**.

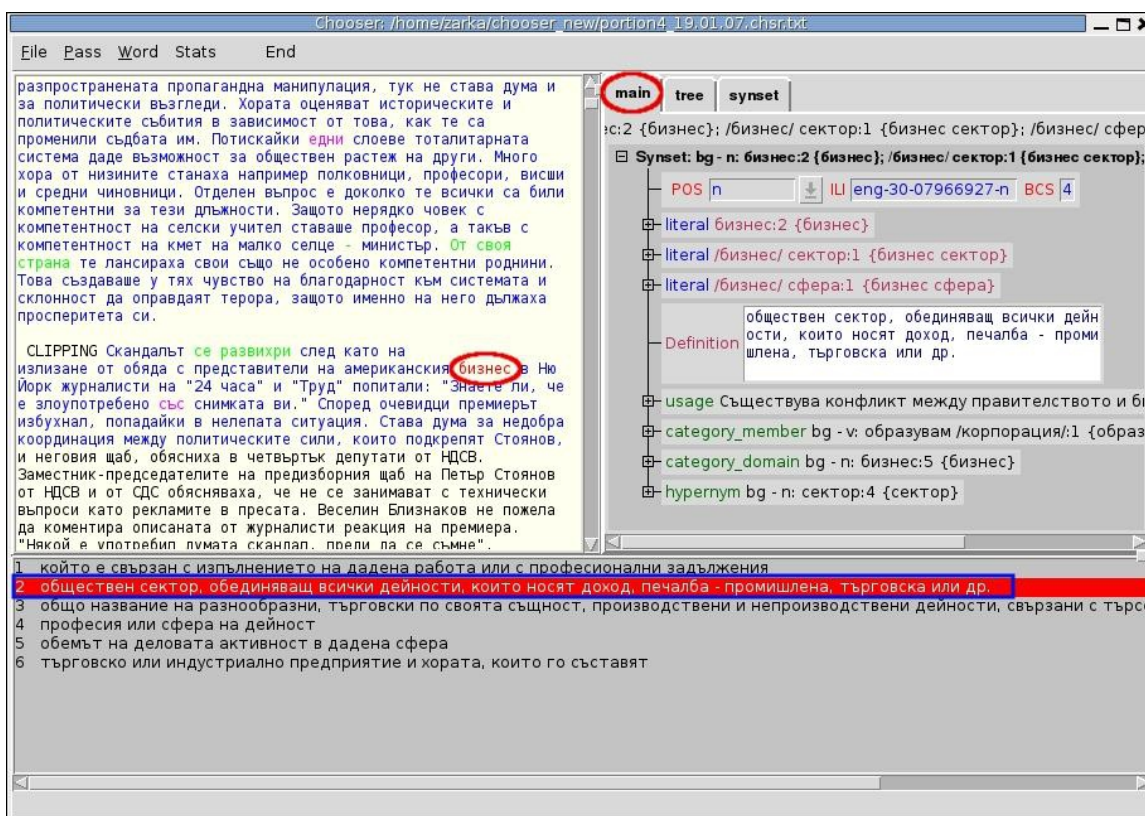


Fig. 2. The Text view showing the currently selected LU *бизнес* (coloured in red)

3.1.3. Info view

The Info view displays portions of the information available in the Wordnet database for a currently selected item in the **List view**. The programme embeds the three display

views of the system for wordnet development Hydra (right pane) - **Main view**, **Tree view** and **Synset view**.

The synchronisation between the **List view** and the **Info view** enables the annotators to make their choices based on a detailed inspection of all the available information associated with a synset.

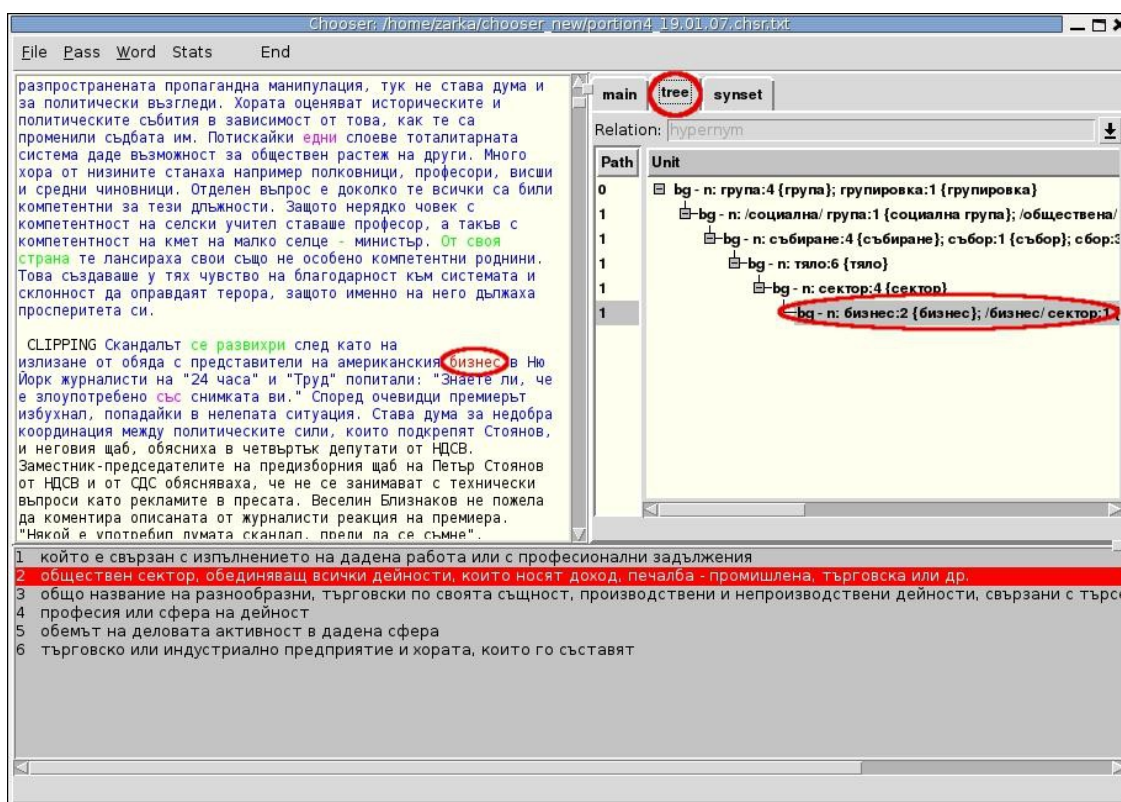


Fig. 3. The Info view displaying the Hydra Tree view mode

A detailed description of Hydra is available in the Hydra user manual:

<http://dcl.bas.bg/Tools/Hydra/Hydra-UserManual.pdf>

3.2. Corpus annotation and editing

3.2.1. Loading and saving files

To load a corpus file:

- (1) use Open from the **File** menu. A browsable dialog appears on the screen.
- (2) browse the file you wish to load and select it.

Once a corpus file is opened, it is displayed in the Text view.

The system saves the file automatically. To explicitly perform the save operation, use the **Save** button in the **File** menu.

To change the name and/or location of the file, use **Save as**.

3.2.2. Selecting LUs

To select/traverse the words in the corpus file, use either of the following actions:

- (1) use the **Left/Right Arrow** keys.

Unless another option is selected from the **Pass** menu, the **Arrow** keys perform linear pass selecting one LU at a time to the left or to the right, respectively.

(2) left-click on a particular word in the Text view. This command allows the user to select an arbitrary word in the text.

For the pass strategies confer the relevant section below (**Traversing the corpus**).

When a language unit is selected, the user is able to view the corresponding annotation candidates, annotate and edit/delete the LU, insert other words with respect to it, define search and traverse operations on it. The possible pass options for a particular LU are: traversing all the instances of the LU by checking **Current word** in the Pass

menu, or traversing all the instances of the particular sense of the LU by checking the **Current sense** box.

Grouping words in multiword expressions

To group two or more words in a multiword expression:

(1) select the individual words that form the MWE by left-clicking on each of them one after the other while holding the **Ctrl** key.

In order for a MWE to be assigned an appropriate lemma and consequently - to be identified in the Wordnet database, the individual words must be grouped in the order in which they appear in the lemma of the corresponding literal in wordnet.

The individual words as part of a MWE may not be in their citation form. The relevant wordform must be typed in the MWE's lemma field. Consider the compound noun **електронна поща** (electronic mail). The adjective **електронен** (electronic) (lemmatised in the same way) agrees in gender with the feminine noun **поща** (mail), and is hence in the feminine singular. In order for the LU **електронна поща** in the corpus to be mapped to the corresponding synset {електронна поща:1; имейл:1; e-mail:1; email:1} ({electronic mail:1; e-mail:3; email:3}) it has to be assigned the correct lemma – **електронна поща**.

To this end the user needs to edit the MWE's lemma as shown in Fig. 6. For details see the section on **Edit operations**.



- The grouping of non-contiguous MWE components and word order variants of MWEs is performed in the same way.
- To ungroup a MWE, left-click on the individual words, while holding the Ctrl key.
- Ellipted components of MWE need to be restored in the MWE lemma.

Consider the following example: **red and white wines**

Select **white wines** as a MWE. It receives the lemma **white wine** and is mapped to the corresponding synset {white wine:1}. In order to be able to map **red** to {red wine:1} in the Wordnet database, select **red** and expand the lemma to **red wine** (the lemmas are represented in the brackets): **red{red wine} and white wines{white wine}**

3.2.3. Annotation

When a LU is selected in the text by means of the **Left/Right Arrow keys** or mouse pointer click, a list of the available annotation options is displayed in the **List view** (Fig. 2).

The annotation of a LU is performed by:

(1) selecting the appropriate sense from the **List view** by (i) browsing the list with the **Up/Down Arrow keys**, (ii) pressing the number or letter key corresponding to the number/letter preceding the relevant definition in the **List view**, (iii) mouse pointer click on the selected item in the list;

(2) pressing **Enter**. The respective definition in the **List view** is highlighted, the LU in the **Text view** is coloured in the relevant colour.



- The users are advised to consult the additional information in the **Info view** - other synonyms in the synset, usage examples, the relations pertaining to the synset - before performing the annotation. The sense distinctions in wordnet may be very fine-grained, which sometimes necessitates close inspection of similar senses.

- There is no specific operation to override the annotation operation. Instead, the user needs to select another sense.
- Changes made to the Wordnet database in Hydra are updated immediately in Chooser. To view the changes concerning a currently selected LU, refresh the **List view** by first disselecting and then re-selecting the particular LU. This is performed by jumping to a previous or next word using the **Left/Right Arrow keys**, or by clicking on any other word and then clicking again on the LU under consideration.

3.2.4. Traversing the corpus

Chooser supports several strategies for traversing the corpus (Fig. 4).

(i) The default option is a linear pass of all the LUs (option **All**). To choose another strategy, check the respective option in the **Pass** menu.

(ii) unannotated LUs (option **Not Chosen**) - traverses the LUs that have not been assigned a sense yet;

(iii) ambiguous LUs (option **Ambiguous**) - passes the LUs to which more than one wordnet sense corresponds

(iv) LUs removed from the Wordnet database (**Phantom**) - traverses those LUs mapped to literals that have been removed from the Wordnet database after the LU has been annotated.

(v) sentence endings (**Sentence End**). This pass option has been defined for the purposes of manual validation of sentence-splitting;

The above options can be combined with the **Current word** pass thus enabling the

traversing of the instances of a current LU that meet the selected option.

A user may also traverse the instances of the particular sense assigned to the current word. Traversing is performed using the cursor movement (**Left/Right Arrow keys**) of the keyboard.

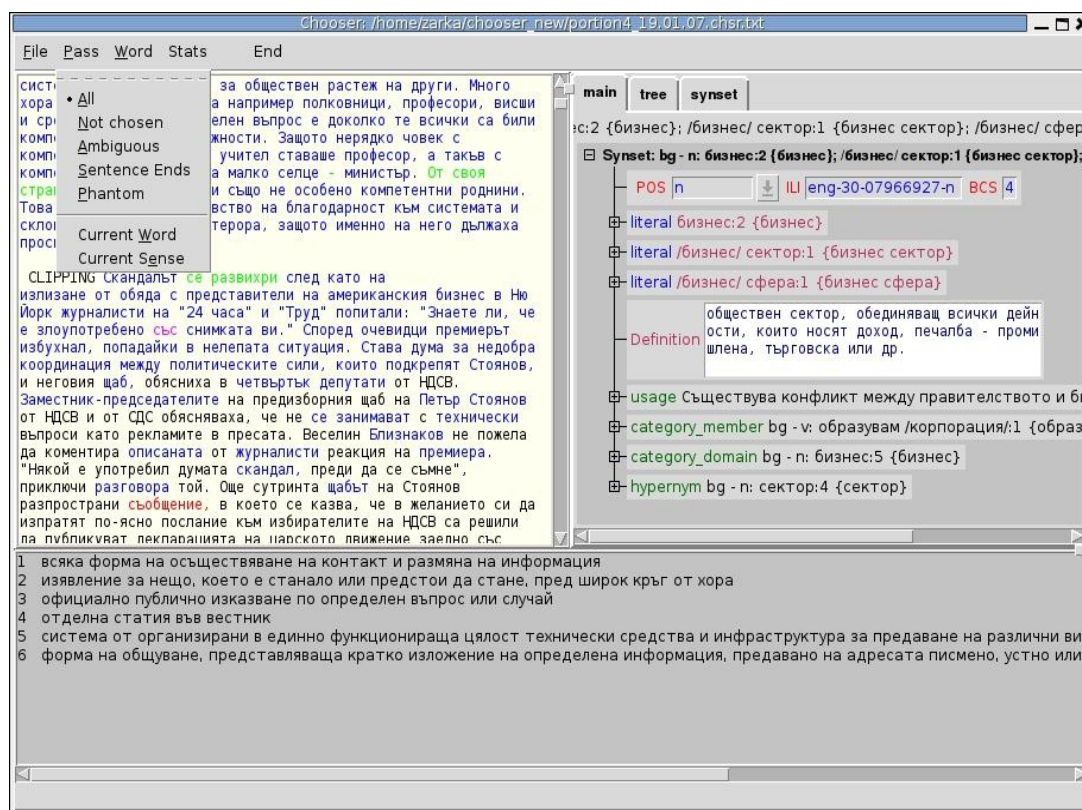


Fig. 4. The Pass menu

3.2.5. Corpus searching and editing

Chooser supports common search operations and two types of operations that involve making changes to the corpus content - **edit** and **add/remove**. The options are selected

from the **Word** menu (Fig. 5).

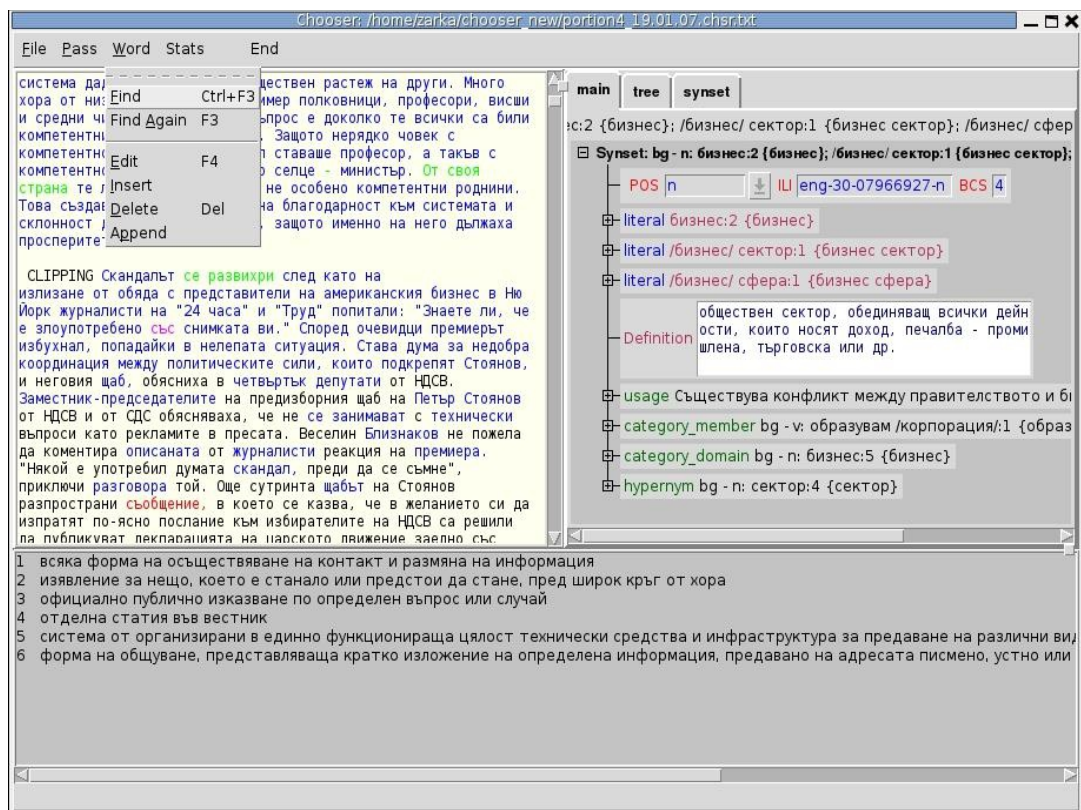


Fig. 5. Chooser's Word menu

3.2.5.1. Edit operations

Edit affords modifications of the wordforms and lemmas of the LUs in the corpus. In such a way misspellings, typos and wrong lemmatisation are corrected directly in Chooser's interface in parallel with the annotation and updated immediately.

To edit a LU:

- (1) press **Edit** in the **Word** menu.

A popup dialog with two fields - **form** and **lemma** appears on the screen (Fig. 6).

- (2) To edit the wordform, retype/correct the form in the **form** field.
- (3) To edit the lemma, retype/correct the form in the **lemma** field.
- (4) Press **ok** to accept the change, or **cancel** to reject it.

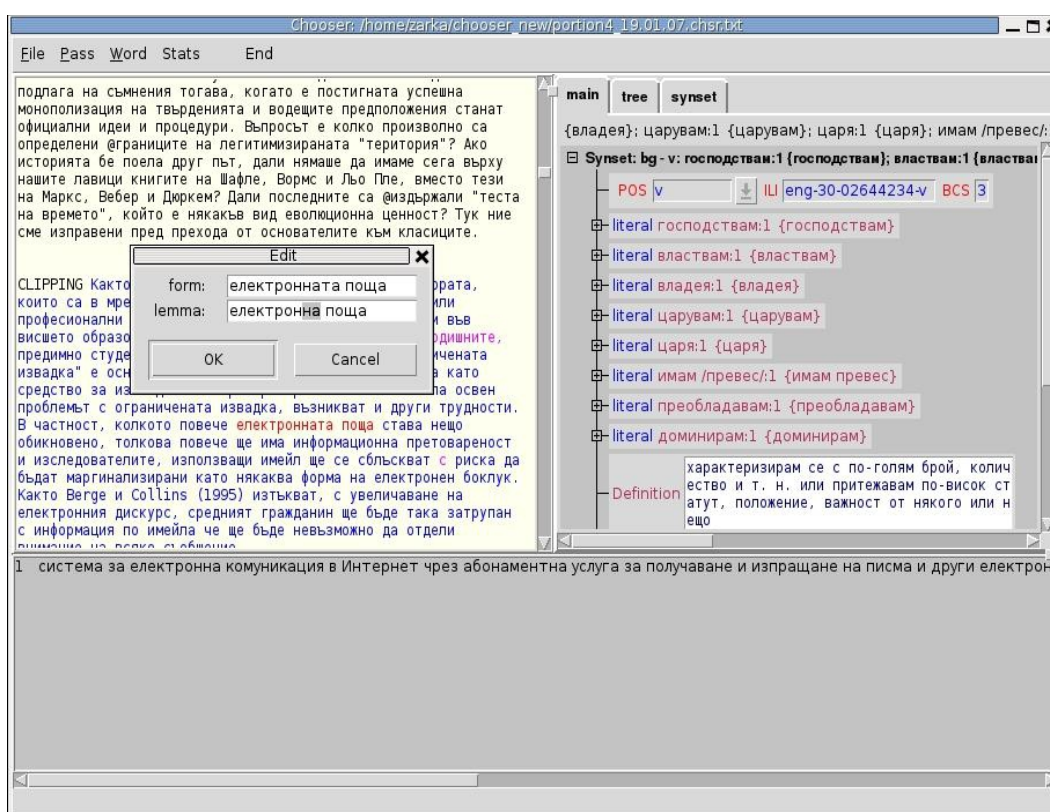


Fig. 6. Chooser's Edit dialog.

The example on Fig 6. shows the Edit dialog with the lemma of the component *електронна* (electronic, feminine singular) corrected from *електронен* (electronic, masculine singular, used as citation form) to form the correct MWE lemma *електронна*

nowa (electronic mail).



- The lemma of a MWE is also corrected in the **Edit** dialog. To do that:

(1) after grouping the elements of the MWE into a single LU (on how to do that see the relevant section above), select it using the **Left/Right Arrow keys** or a mouse pointer click.

(2) Select **Edit** from the **Word** menu.

(3) If needed, correct the lemmas of the elements in the **lemma** field, so that they are in the appropriate form.

(4) Press **ok**. The definition of the relevant synset appears in the **List view**.

- If the constituents are ungrouped they will restore their original lemmas.

Useful tips

To add a punctuation mark in the corpus text, append the relevant symbol to the wordform (in the **form** field and not in the **lemma** field) of the word it should be attached to.

Capital letters marking the beginning of a sentence are also inserted in the **form** field and not in the **lemma** field. For instance, the lemma of a common noun such as **bus** is **bus**. However if the word **bus** begins a sentence, it should be capitalised (in the **form** field).

If the lemma contains capital(s), they should be accounted for in the **lemma** field, e.g. the lemma of a proper noun such as **New York** should be **New York**.

3.2.5.2. Adding/deleting words

Insert

To insert a word:

- (1) select **Insert** in the **Word** menu. An **Edit** dialog appears on the screen.
- (2) Type the wordform you want to add into the text and the respective lemma.
- (3) Press **ok** to accept, or cancel to **reject** the operation.



- The new word is inserted before the current LU.

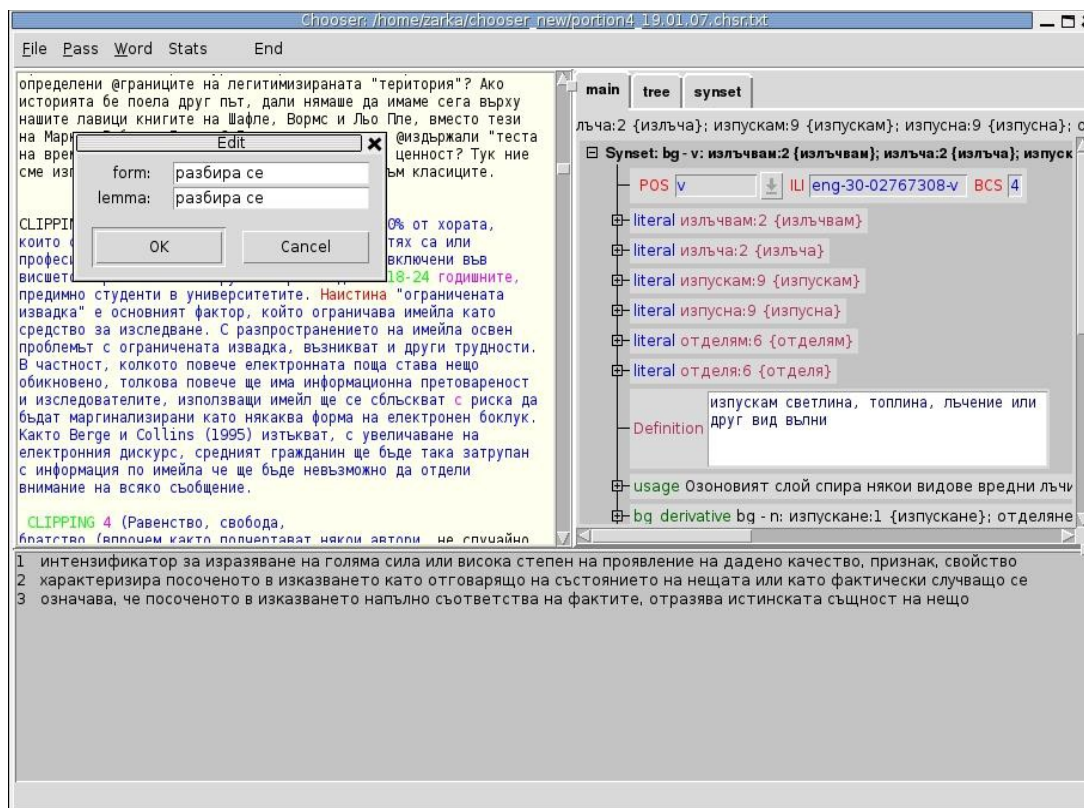


Fig. 7. The language unit *разбира се* (of course) is just added in the dialog, but not yet confirmed.

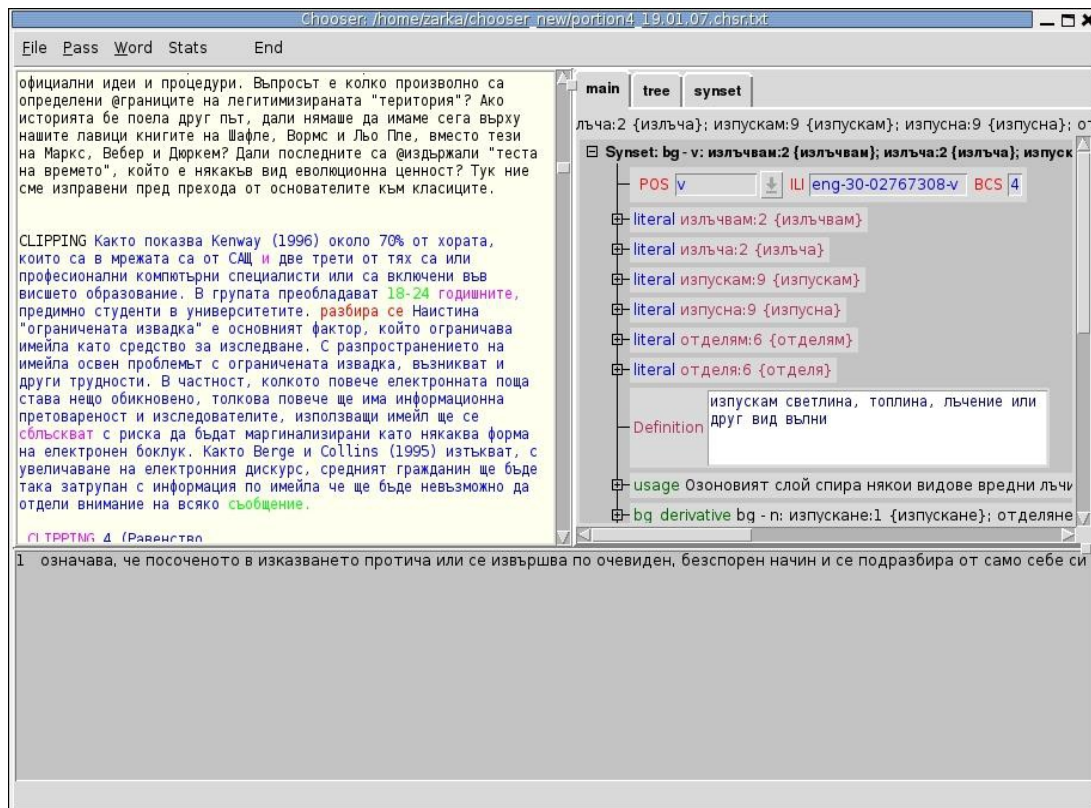


Fig. 8. After the insertion is confirmed, the freshly added LU *разбуца се* appears in the text.

Append

Append is used to add a word to the end of a file. To append a word, follow the steps described for the **Insert** operation.

Delete

To delete a LU:

(1) select **Delete** from the **Word** menu or press **Delete** on the keyboard. A popup dialog asking to confirm or cancel the operation appears on the screen.

(2) Press either of the two options to confirm or cancel.

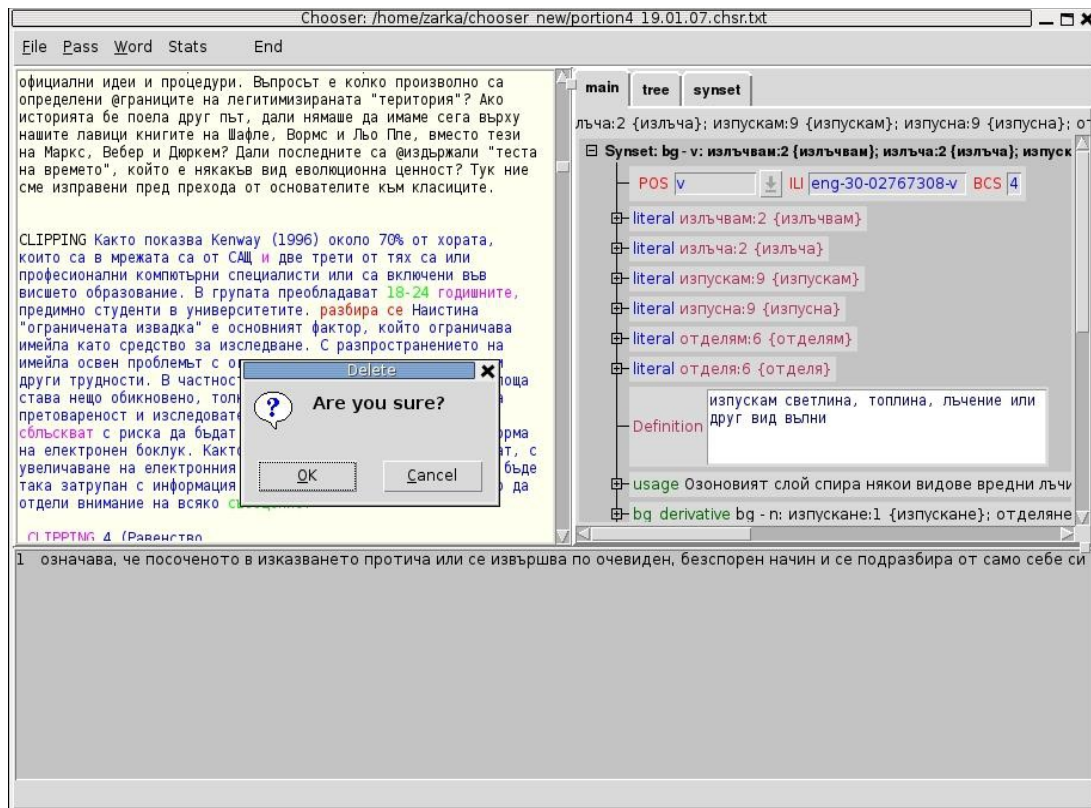


Fig. 9. Delete dialog.

Useful tips

In the corpus two or more individual tokens may be attached to each other (without a separator) as a result of wrong tokenisation - for instance, a number - to a following word, a special symbol %, \$, etc. - to a previous word, etc. Besides one-word compounds may be typed as separate words, and as a result - constitute more than one token, and vice versa - multiword expressions may be typed as a single token.

To split two (or more tokens), use the **Insert** and **Edit** functions:

- (1) insert an additional token, type the relevant wordform and lemma
- (2) edit the original token's wordform and lemma

To merge two (or more tokens), use the **Delete** and **Edit** functions:

- (1) delete any of the tokens
- (2) edit the other's lemma and wordform.

3.2.5.3. Search

Chooser supports simple and regular expression (employing Python's regular expression syntax) search functions over:

- (i) wordforms
- (ii) lemmas

Case sensitive search and search direction (forward/backward) are selected from the menu.

To enable the search function:

- (1) select the option **Search** from the **Word** menu. A Search dialog pops up.
- (2) type a word in the **word** field.
- (3) to search for this particular form in the corpus, check the **form** box
- (4) to search for all the wordforms of a lemma, check the **lemma** box
- (5) check any of the other possible criteria (search direction and case sensitivity).
- (6) press **ok** to initiate the search, or **cancel** to disable it.

(7) to resume the search after a match is found, use the **F3 + Ctrl key**. The default search direction is forward. **Search backwards** may be selected from the search dialog (**F3**).

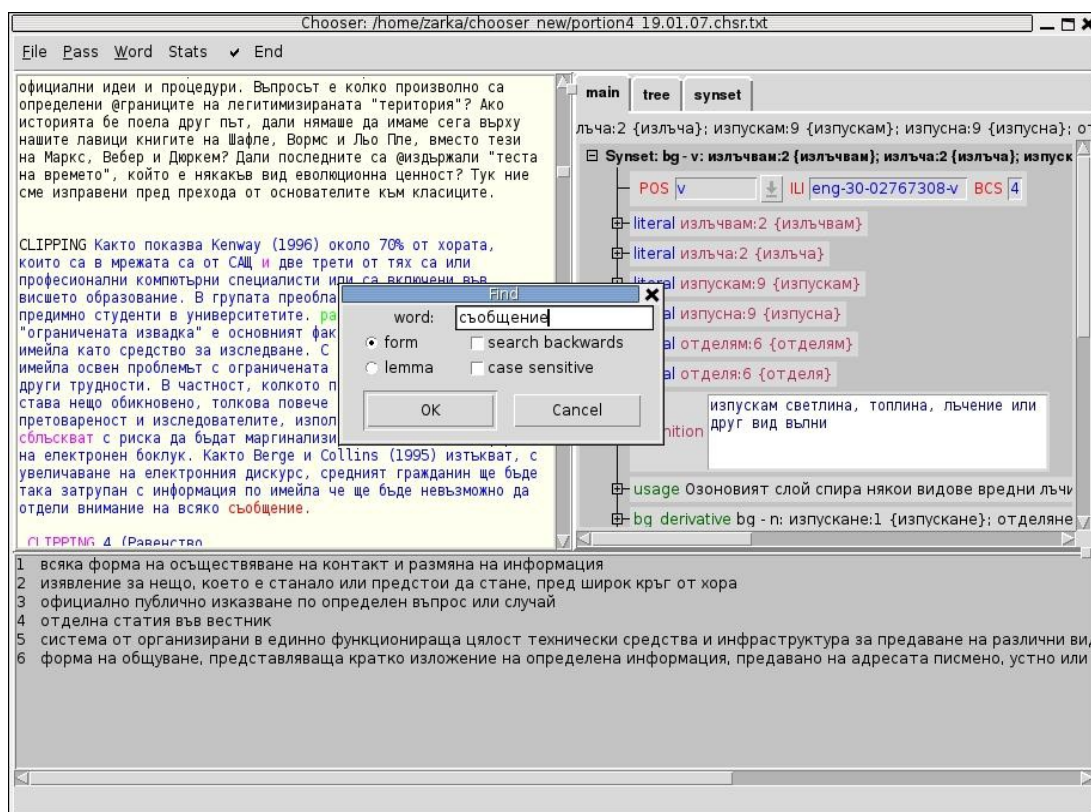


Fig. 10. Chooser's search dialog

References

Koeva, S., S. Leseva, B. Rizov, E. Tarpomanova, T. Dimitrova, H. Kukova, M. Todorova. Design and Development of the Bulgarian Sense-Annotated Corpus. In Proceedings of

the Third International Corpus Linguistics Conference (CILC), 7-9 April 2011, Valencia, Spain.

Koeva, S., B. Rizov, S. Leseva. Chooser - a Multitask Annotation Tool. In Proceedings of the 6th Language Resources and Evaluation Conference, Marrakech, Morocco, 28-30 May, 2008. ISBN 2-9517408-4-0.