

# COREX: A tool for exploiting the Corpus Gesproken Nederlands (CGN)

## Manual

Authors: Birgit Hellwig and Erik Weijers

© 2004, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

Update in January 2006 by Griet Depoorter, TST-centrale, INL

### **Introduction**

The Corpus Gesproken Nederlands (CGN) is a database of recordings and annotations that contains almost nine million spoken Dutch words. The Corex program allows you to listen to the speech files, to view the multiple annotations and to conduct queries in the corpus. It supports the following features:

- easy navigation through subcorpora, either based on predefined or user-defined metadata groupings such as the speaker's sex and age, the region in which (s)he grew up, the text type and other metadata.
- playing of speech signal, synchronised with annotation data
- display of, search through and statistical analysis of annotation data
- display and search of metadata descriptions (e.g. information about the speakers, information about socio-environmental settings of the recording).

This manual describes the Corex program commands. It is organised in a sequential manner based on the windows and panels that open up in the program. The following is a list of the Corex windows and panels and their corresponding sections in this manual:

- 1 Corpus Browser Window
- 2 Metadata Search Panel
- 3 Corex Viewer
- 4 Content Search Panel
- 5 Statistics Panel
- 6 Lexicon Tool Panel
- 7 Syntax Search Panel

## **Errata**

---

The following terminology has been changed in the metadata but not yet in this manual:

Participant → Actor

Participants → Actors

AnnotationUnit → WrittenResource

The new terminology will be used in the next version of this manual.

TST-centrale, 24 January 2006

# **Table of contents**

<b>1</b>	<b>Corpus Browser Window</b>	<b>5</b>
1.1	Metadata Descriptions Tree Panel	6
1.1.1	Navigating in the CGN corpus	6
1.1.2	Displaying information	8
1.1.3	Selecting parts of the corpus for purposes of analysis	11
1.2	Bookmarks Panel	14
1.3	Info/Content Panel	15
1.4	Description Panel	17
1.5	Menu	17
1.5.1	The Search menu	17
1.5.2	The Options menu	18
1.5.3	Help menu	19
1.5.4	The file menu	19
<b>2</b>	<b>Metadata Search Panel</b>	<b>20</b>
2.1	Specify the search options	21
2.1.1	Select the category to be searched	21
2.1	Initiate and stop the search	26
2.2	Display the search results	26
2.3	Save the search results	26
<b>3</b>	<b>Corex Viewer</b>	<b>29</b>
3.1	File menu	30
3.1.1	Print view	30
3.1.2	Print all	30
3.2	Panel menu	31
3.2.1	Export to HTML	31
3.3	Options menu	31
3.3.1	Show Metadata	31
3.3.2	Play only segment	31
3.3.3	Time sync	32
3.3.4	Visible tracks	32
3.3.5	Preferences	33
3.4	Tools menu	34
3.4.1	Praat Synch	34
3.5	Audio menu	34
3.5.1	Audio Player	34
3.5.2	Waveform Panel	36
<b>4</b>	<b>Content Search Panel</b>	<b>37</b>
4.1	Track Selection Menu and Text Field Box	38
4.1.1	Phonetics	40
4.1.2	Prosodic search	42
4.1.3	Marked Words	44
4.2	Regular Expression Search	44
4.3	Case-sensitive search	46
4.4	Add new constraint button	46
4.4.1	Performing a query within results	48
4.5	Delete last constraint button	49
4.6	Save Result	49
4.7	Read result	50
4.8	Save results as corpus	51
<b>5</b>	<b>Statistics Panel</b>	<b>53</b>
<b>6</b>	<b>The Lexicon Tool Panel</b>	<b>55</b>
6.1	Performing a Query in the Single Word Lexicon	55
6.1.1	Column visibility settings	57
6.2	Searching Within Results	57
6.3	Saving and printing the search results	58
6.4	Searching the CGN with word/lemma ID	58
6.5	Statistics based on word/lemma ID	60
6.6	The multi-word lexicon	61

<b>7</b>	<b>Syntax Search Panel .....</b>	<b>62</b>
7.1	The TIGERGraph Viewer.....	62
7.2	Syntax Search .....	64
7.2.1	Specify the search options .....	64
7.2.2	Saving the search results.....	65
	<b>Appendix A: example of .res file .....</b>	<b>66</b>
	<b>Appendix B: CGN Metadata .....</b>	<b>67</b>

# 1 Corpus Browser Window

Starting the Corex program opens up the Corpus Browser window.

In the **Corpus Browser** window you can view and access the Corpus Gesproken Nederlands (CGN): you can read information about the kind of data that is contained in the corpus, you can access the annotation and audio files, and you can initiate searches and do statistical counts.

The **Corpus Browser** window contains the following four panels:

1.1 Metadata Descriptions Tree panel

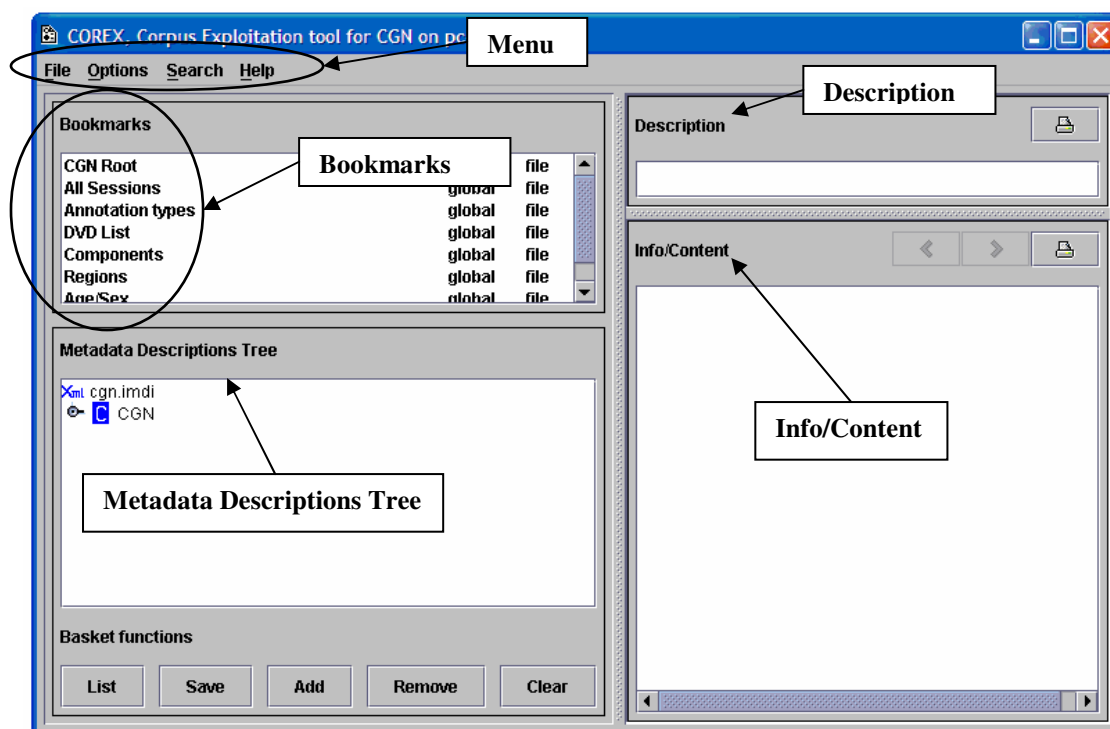
1.2 Bookmarks panel

1.3 Info/Content panel

1.4 Description panel

Besides these panels, there is the main menu:

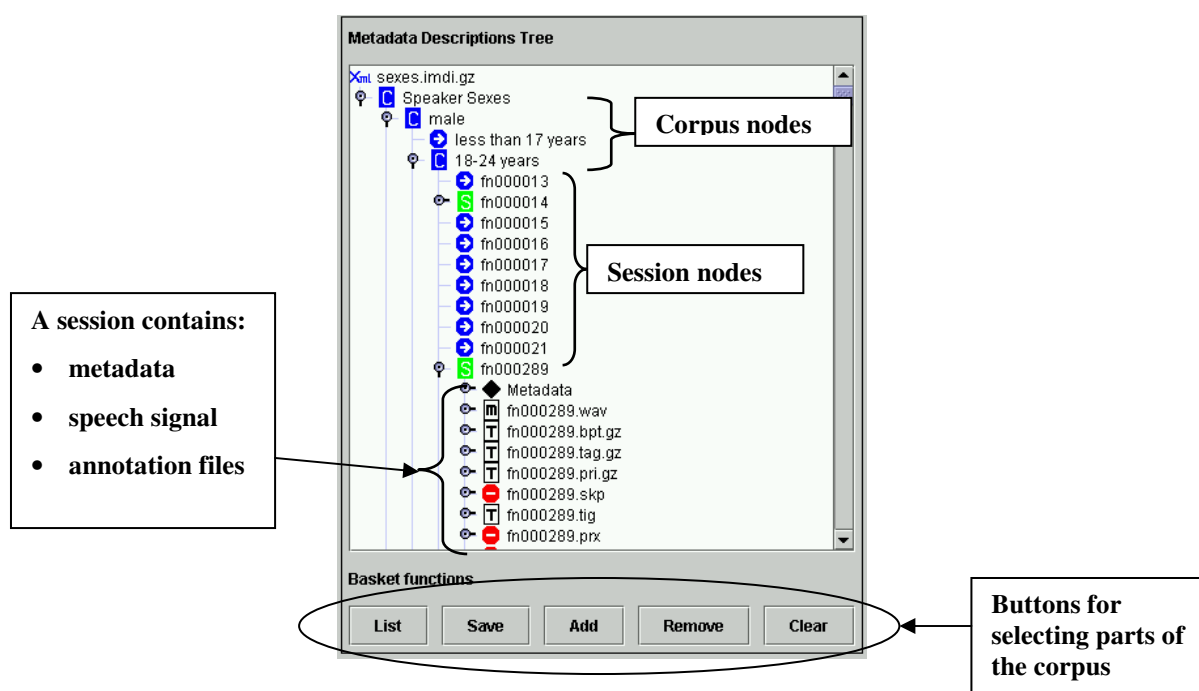
1.5 Menu



## 1.1 Metadata Descriptions Tree Panel

The CGN corpus is a collection of 12780 recordings and corresponding annotation data. The **Metadata Descriptions Tree** panel allows you to navigate in a predefined structure of the Corpus Gesproken Nederlands (CGN). It gives access to the so-called *sessions*, which are the basic elements of the CGN structure. A session includes the audio file and the corresponding annotation files and metadata.

At the bottom of the panel, there are five buttons that allow you to select parts of the CGN corpus for purpose of analysis.



### 1.1.1 Navigating in the CGN corpus

The top node of the CGN is displayed as the **Corpus Browser** window initially opens up. Double-click on the CGN icon to display its corpus nodes; double-click on any corpus node to open it and to display the next level in the hierarchy.

**! Note:** The **Metadata Descriptions Tree** distinguishes between open and closed corpus and session nodes. These are represented by different icons:



Icon of an open corpus/session.



Icon of a closed corpus/session.

Most of the program commands do not work when the node is closed. Therefore, if any of the commands do not seem to work, please make sure that the node is open. Double-click on its icon to open it.

! Note: Because of the large amount of data that is loaded, it may take some time until Corex responds to the ‘open’ command.

The CGN corpus branches into multiple subcorpora. Subcorpora can be partly overlapping. They just provide a different perspective on the same data.

- **All:** this is simply a list of all sessions, ordered by number.
- **DVDs:** the CGN corpus as it is grouped according to the audio DVDs that are supplied with Corex.
- **Annotation types:** Subcorpora that contain annotation types (such as prosody) that are provided only for *part* of the sessions.
- **Components:** the CGN corpus as it is grouped according to text type (e.g. spontaneous conversations).
- **Regions:** the CGN corpus as it is grouped according to the region where the speaker(s) lived between the ages of four and sixteen.
- **Speaker sexes:** the CGN corpus as it is grouped according to the sex of the speaker(s).
- **Speaker ages:** the CGN corpus as it is grouped according to the age of the speaker(s).

The lowest level in the CGN structure is the session. Each session contains the audio data and is linked to the corresponding annotation files as well as the metadata. E.g., we can open the session labelled “*fn000001*” and see in the metadata that it contains a live commentary (text type *I*) of a 26 years old male speaker from the region Gelders Rivierengebied (Region N2c). See section [1.3](#) for the procedure of opening a session. See Appendix B for an overview of all metadata codes.


Session names that start with ‘fn’ are Dutch, session names that start with ‘fv’ are Flemish.

Each session node contains the following kind of information:

- metadata descriptions, i.e. information *about* the speaker(s), the recording and annotation data (see sections [1.3](#) and [1.4](#)),
- links to a **\*.wav** file (the audio file),
- links to a number of annotation files (**.bpt**, **.tag**, **.pri**, **.skp**, **.tig**, **.prx**).
  - The **.bpt** file contains the verified broad phonetic transcription. This was done only for a part of the CGN.
  - The **.tag** file contains amongst others the part-of-speech tags (POS tags).
  - The **.pri** file contains the orthographic transcription.
  - The **.skp** file contains information about the linkage between the audio signal and the orthographic transcription. The linkage is on word level. This means that the time codes for the beginning and end of each word are provided. These time codes were set manually for about ten percent of the corpus. The remaining part was generated automatically. In case the manually performed transcription is available for a session, it ‘overrides’ the automatically performed transcription.

- The **.tig** file can be read using the Tiger program for syntactical data (this is not a Corex tool, but it is integrated in Corex).
- The **.prx** file contains the prosodic annotation. For about five percent of the sessions there are two prosodic transcriptions, according to the two schools of annotating prosody that exist in the Netherlands and in Flanders (this makes a total of four schools. However, since each session is either Dutch or Flemish, a session can only have two **.prx** files).

Double-click on any session node to open a viewer panel (the **Corex Viewer**), which displays the session's annotations (see section 3). Each session has a corresponding audio file. In order to use the Corex **Audio Player** or **Waveform Panel** features, you need the DVD that contains the relevant audio file (see sections 3.5.1 and 3.5.2).

! Note: You can access the **Corex Viewer** and the **Audio Player** and **Waveform Panel** by double-clicking on the session icon:  or by right-clicking on an opened session and choosing the option **Corex Viewer**.

Clicking on any of the annotation files will *not* start the **Corex Viewer**. Nor will clicking on the audio file start the **Audio Player**.

! Note: Not all annotation types are available for every session. Available and non-available files are represented by different icons:



file is not available



available transcription file.



available **.wav** file

! Note that if a **.wav** (speech) file is not available, you will be prompted to provide a DVD with audio files.

### 1.1.2 Displaying information

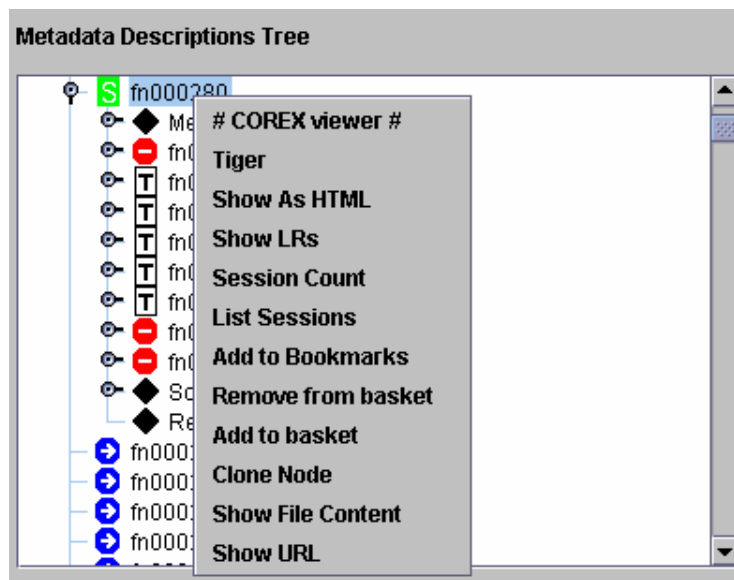
A left mouse click on any node in the **Metadata Descriptions Tree** panel will highlight this node; and a right mouse click on any open highlighted node will display a drop-down menu, offering you a number of different options.

! Note: Not all options are available for all nodes.

! Note: For some nodes, one of the options is marked with the symbol #. In this case, the corresponding option will start automatically whenever you double-click on the open node.



- The following options are available if you right click on the highlighted (green) *session node*; in the below example this is session fn000280:



COREX viewer	Opens the <b>Corex Viewer</b> , which displays the session's annotations and gives you access to the audio data (see section <a href="#">3</a> ). This is the default setting for any session - i.e. double-clicking on the session node will automatically open the <b>Corex Viewer</b> .
Tiger	Opens the <b>TIGERGraph Viewer</b> (see section <a href="#">7</a> ).
Show as HTML	Displays the metadata neatly formatted for printing.
Show LR's	Displays the information for all annotation and audio files contained under this node in the <b>Info/Content</b> panel.
Session Count	Displays the number of sessions contained under this node in the <b>Info/Content</b> panel.
List Sessions	Lists all sessions contained under this node in the <b>Info/Content</b> panel.
Add to Bookmarks	Adds the node to the <b>Bookmarks</b> panel.
Remove from basket	Removes the node from the list of nodes to be analysed (see section <a href="#">1.1.3</a> ). This option is identical to the <b>Remove</b> button in the panel of basket functions below the <b>Metadata Descriptions Tree</b> panel.

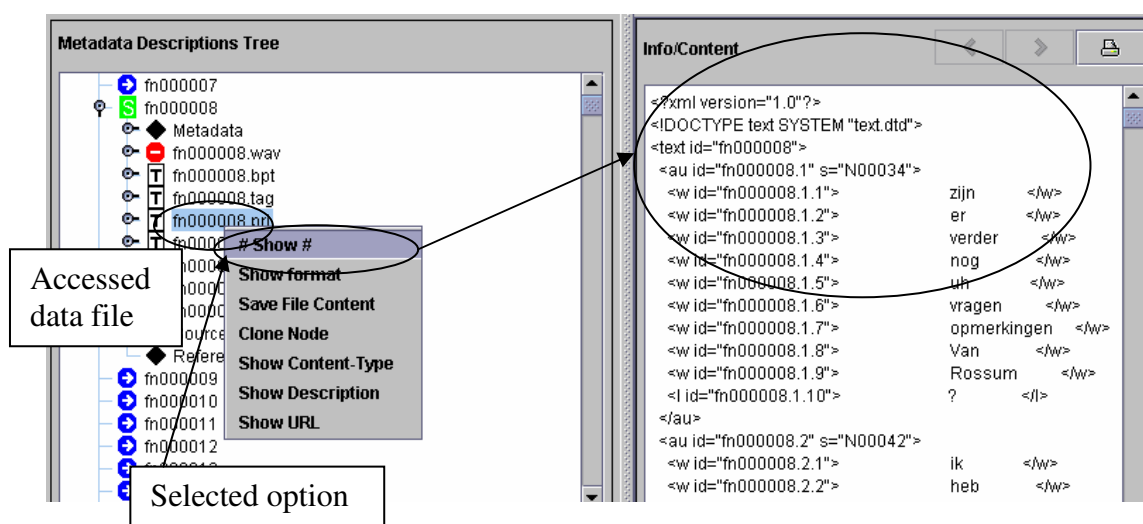


Show Content-Type	Displays the file format in the <b>Description</b> panel.
Show Description	Displays the file description in the <b>Description</b> panel.
Show URL	Displays the directory information for the file in the <b>Description</b> panel.

- The following additional option is available only for result corpora (subcorpora you have composed yourself) in the SEARCH RESULTS:

Remove	Removes the node from the <b>Metadata Descriptions Tree</b> panel. Note: After selecting <b>Remove</b> , the icon of the removed node remains visible although it cannot be accessed anymore. The icon will be gone after exiting and re-entering the node <b>Search Results</b> .
--------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

The preferred way of displaying the annotation data is by use of the Corex Viewer, because in that way the annotations are linked to the speech signal (see section 3). It is also possible, however, to display all information either in the **Info/Content** panel (see section 1.3) or in the **Description** panel (see section 1.4). The following illustration is an example of how selected information is displayed in the **Info/Content** panel:



### 1.1.3 Selecting parts of the corpus for purposes of analysis

The Corex program allows you to do metadata and content searches as well as statistical counts in the corpus (see sections 2, 4 and 5). By default, the analysis is done throughout the whole corpus. However, it is possible to limit the analysis to a part of the corpus: to one (or several) corpus and/or session node(s). The five buttons

at the bottom of the **Metadata Descriptions Tree** panel are used for the selection process.

To select a corpus or session node, do the following:

1. Open the node by double-clicking on it.
2. Highlight the node by clicking on it with the left mouse button.
3. Either click the **Add** button at the bottom of the **Metadata Descriptions Tree** panel.

Or click with the right mouse button on the highlighted item, and select **Add to basket** from the drop-down menu.

The icon of any selected node will change its colour to grey, e.g.:



Non-selected node.



Selected node.

Once an item is selected, the **List** button will be highlighted in red:



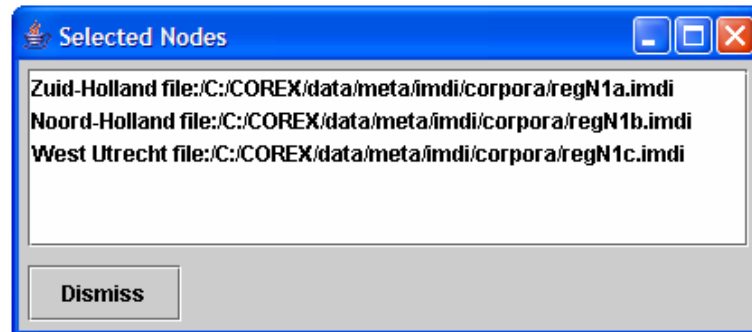
**! Note:** keep in mind that if this button is highlighted in red, any search will only be performed in the subcorpus that is selected. If nothing is in the basket – if the list is empty – any search will be performed in the entire CGN.

4. Repeat this process to add other nodes to your selection.

**! Note:** You can only select a node that is open. Double-click on it to open it.

**! Note:** It is possible to select several different nodes, e.g. the nodes (a) ‘male’ and (b) ‘The Netherlands’. The analysis will then be done over all sessions contained under the node ‘male’ (regardless of region) and all sessions contained under the node ‘The Netherlands’ (regardless of sex). This option does not allow you to limit the search to, e.g., all male speakers that grew up in the Netherlands. To limit your search in such a way, make use of the metadata search options (see section [2](#)).

Click the List button to view a list of all selected nodes, e.g.:



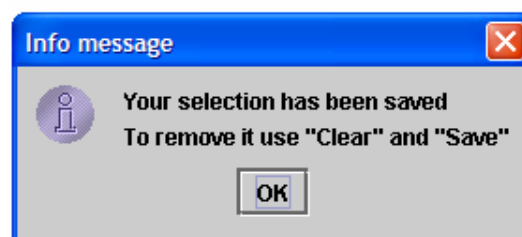
Click the Dismiss button to return to the **Metadata Descriptions Tree** panel.

You can remove all selected nodes from this list by clicking the Clear button at the bottom of the **Metadata Descriptions Tree** panel.

To remove a specific node from the list, do the following:

1. In the **Metadata Descriptions Tree** panel, highlight the node by clicking on it with the left mouse button.
2. Either click the Remove button at the bottom of the panel.  
Or click with the right mouse button on the highlighted item, and select Remove from basket from the drop-down menu.
3. Repeat this process to remove other nodes from your selection.

You can save the selected list for future use. Click the Save button. The following message informs you that your list has been saved:

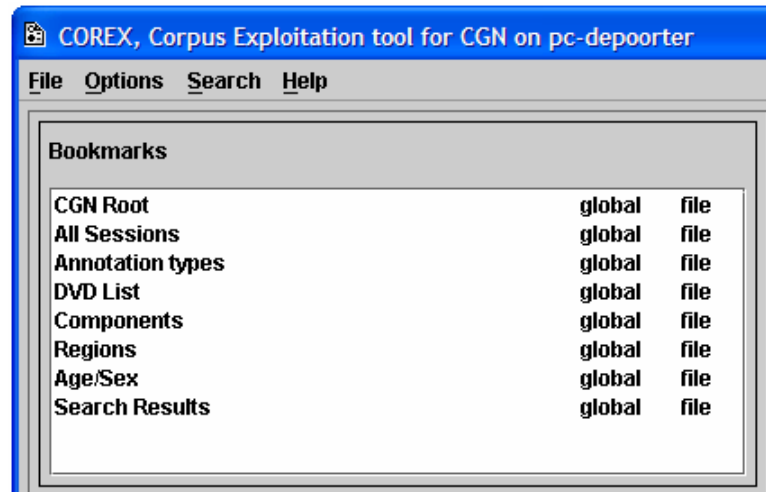


**! Note:** Once you have saved a selected list, you can only remove it by first clicking the Clear button (to remove all selected nodes) and then the Save button (to save an empty list).

When you are satisfied with your selection, turn to the pull-down menu Search in order to initiate searches or statistical counts (see section [1.5](#)).

## 1.2 Bookmarks Panel

The **Bookmarks** panel of the **Corpus Browser** window – visible at the top left part - displays shortcuts to various nodes in the corpus just under the CGN root. Having a bookmark allows you to immediately access such a node, without being obliged to first navigate in the entire corpus. By default, the following bookmarks are displayed:



- **CGN Root:** the entire CGN corpus.
- **All Sessions:** a list of all sessions, sorted by name, no thematic grouping.
- **Annotation types:** contains three subcorpora - phonetic, prosodic and syntactic annotations - that are provided for a minor part of the CGN.
- **DVD List:** the CGN corpus as it is grouped according to the audio DVDs that are supplied with Corex.
- **Components:** the CGN corpus as it is grouped according to the discourse genre.
- **Regions:** the CGN corpus as it is grouped according to the region where the speaker(s) lived between the ages of four and sixteen.
- **Age/Sex:** the CGN corpus as it is grouped according to the sex and age of the speaker(s).
- **Search Results:** the saved result corpora of your searches (see section [4.8](#)).

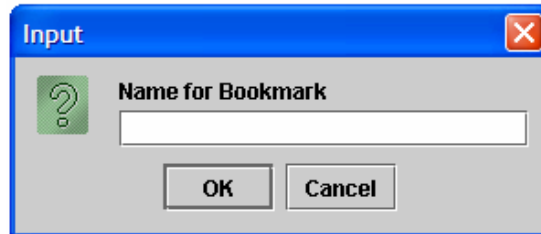
Double-click on any of these items in the **Bookmarks** panel to open the corresponding node in the **Metadata Descriptions Tree** panel.

In addition to the predefined bookmarks, you can create your own bookmarks. Do the following:

- 1 In the **Metadata Descriptions Tree** panel, open the relevant corpus or session node by double-clicking on it.

- 2 Highlight the open node by clicking on it with the left mouse button.
- 3 Right-click on the highlighted node and select **Add to Bookmarks** from the drop-down menu.

The following dialogue box appears:



- 4 Type in a name for the new bookmark. The new bookmark is added to the **Bookmarks** panel.

These bookmarks remain available every time you restart Corex. To remove a bookmark, do the following:

- 1 In the **Bookmarks** panel, click on the bookmark that you want to remove.
- 2 Right-click on that bookmark and select **Delete Bookmark** from the drop-down menu.

The bookmark is deleted without further warning.

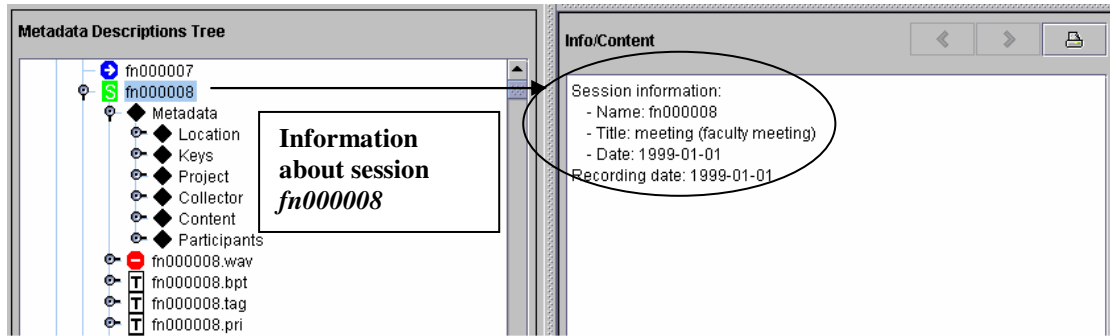
**! Note:** You can only delete those bookmarks that were created by yourself, but never the ones that are predefined by Corex.

### 1.3 Info/Content Panel

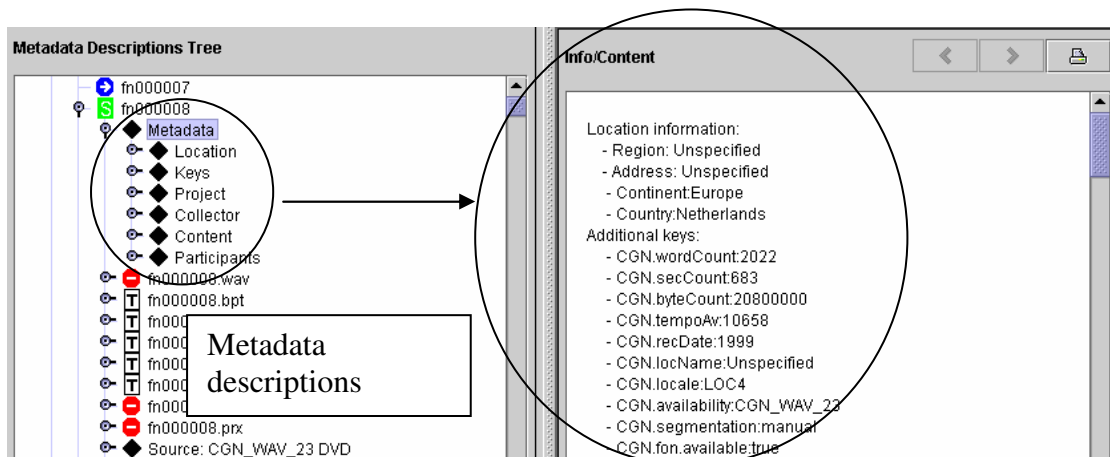
The **Info/Content** panel displays information about corpus nodes, session nodes and files. To read the information, click on the corresponding icon in the **Metadata Descriptions Tree** panel.

**! Note:** The information is only displayed when the corpus or session node is open. Double-click on it to open it.

In the case of corpus and session nodes, information about the name of the node, the title, and (in the case of sessions) the recording date is displayed, e.g.:



In the case of a **Metadata** node, the following kind of information is displayed:

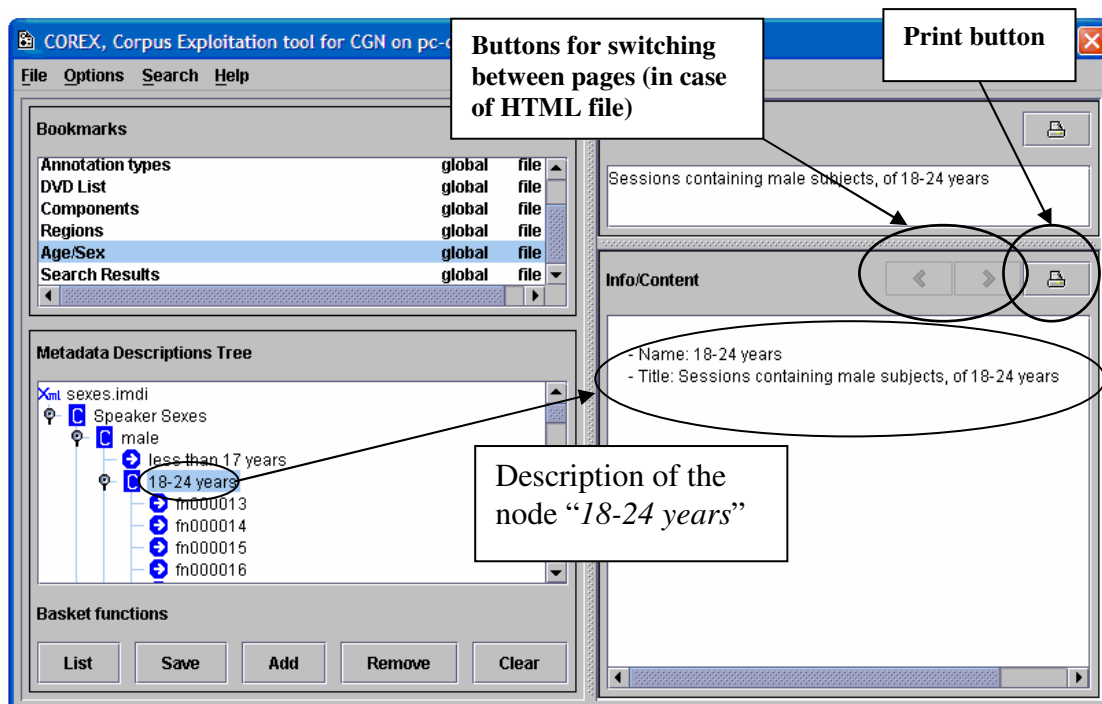


For an overview of all metadata, see Appendix B.

The **Info/Content** panel can optionally display additional information such as a list of all the sessions contained under a node, a count of all the sessions contained there, the content of XML files, the format of file(s), the directory information of file(s) and the directory information of the file containing the metadata descriptions. To view this information, do the following:

1. In the **Metadata Descriptions Tree** panel, click on the node or file of which you want to view the information.
2. Right-click on the highlighted node or file.
3. Select an option from the drop-down menu (see section [1.1.2](#) for the available options). The relevant content is displayed in the **Info/Content** panel.
  - If an HTML page, for example the CGN documentation, is displayed in the **Info/Content** panel you can navigate in this document by clicking the < and > buttons.
  - Printing of the content is possible. Click on the **print** button.





## 1.4 Description Panel

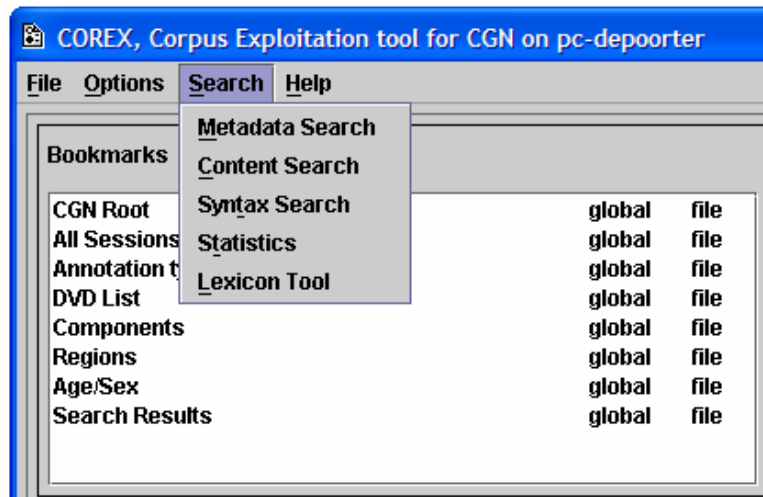
The **Description** panel displays a brief description of corpus nodes, session nodes and files. To read the description, click on the corresponding icon in the **Metadata Descriptions Tree** panel.

! Note: The information is only displayed when the corpus or session node is open. Double-click on it to open it.

## 1.5 Menu

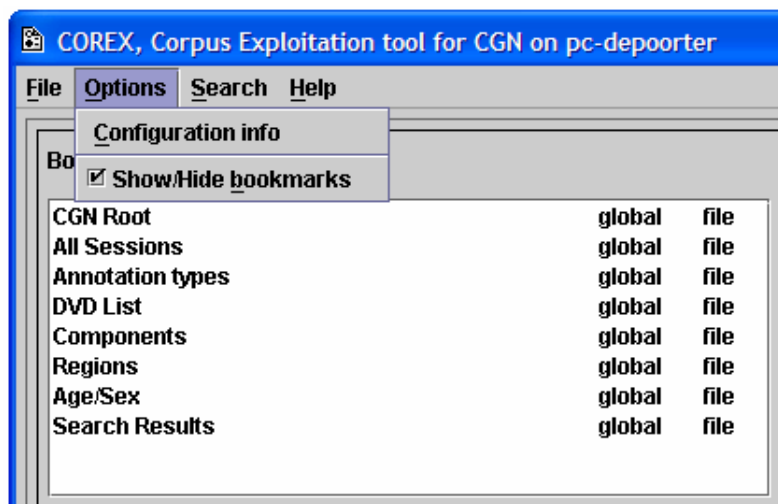
### 1.5.1 The Search menu

This is the most important pull-down menu. From her, we can initiate searches, perform statistical counts in the CGN corpus as well as accessing the lexicon.



- Metadata Search is discussed in chapter 2
- Content Search is discussed in chapter 4.
- Syntax Search is discussed in chapter 7.
- Statistics option is discussed in chapter 5.
- Lexicon Tool is discussed in chapter 6.

### 1.5.2 The Options menu



Click this button for the following two options:

- **Configuration info:** displays information about the configuration, which you might need when reporting a bug to the developers.
- **Show/Hide bookmarks:** offers the possibility to hide the bookmarks panel.

### 1.5.3 Help menu

**About:** click this button to view the copyright and version information.

### 1.5.4 The file menu

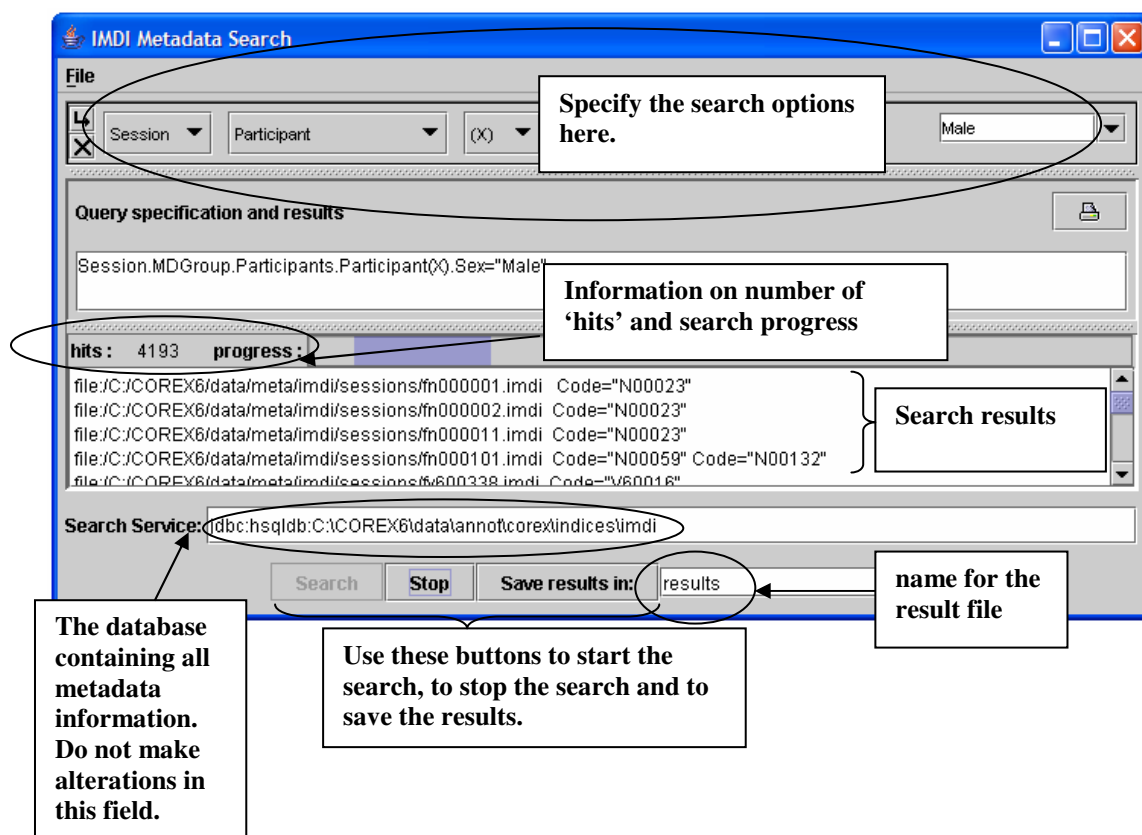
- **Exit:** click this button to exit the Corex program.

**! Note:** When you open a second **Corpus Browser** window (either by the option **Clone Node** (see section [1.1.2](#)), or by double-clicking on search results in the **Metadata Search** panel (see section [2.3](#))), the **Exit** button of the second window turns into a **Close** button. This allows you to close the second window without exiting the Corex program.

## 2 Metadata Search Panel

The **Metadata Search** panel is accessed via the pull-down menu option Search → Metadata Search.

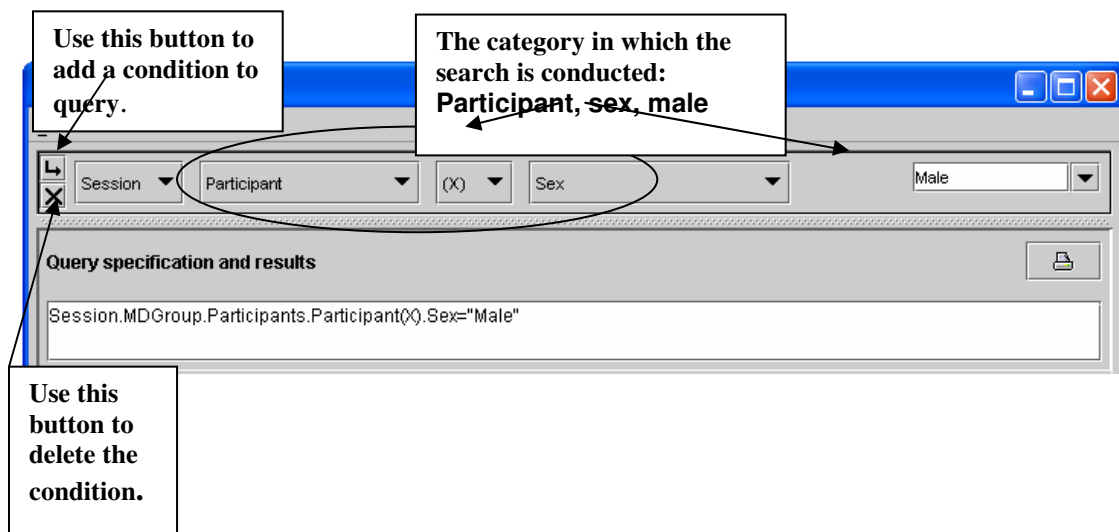
The **Metadata Search** panel allows you to search for sessions that comply with certain metadata conditions (for example, search for all sessions with speakers of a specific age and from a specific region). This can be done throughout the entire Corpus Gesproken Nederlands - the default setting - or throughout selected parts of the corpus (see section [1.1.3](#)).



- Use the Exit button in the File menu to close the **Metadata Search** panel.

## 2.1 Specify the search options

The search options are specified in the topmost part of the **Metadata Search** panel, e.g.:



The following search steps have been taken:

- 2.1.1 Select the category to be searched.
- 2.1.2 Add or delete a search query.

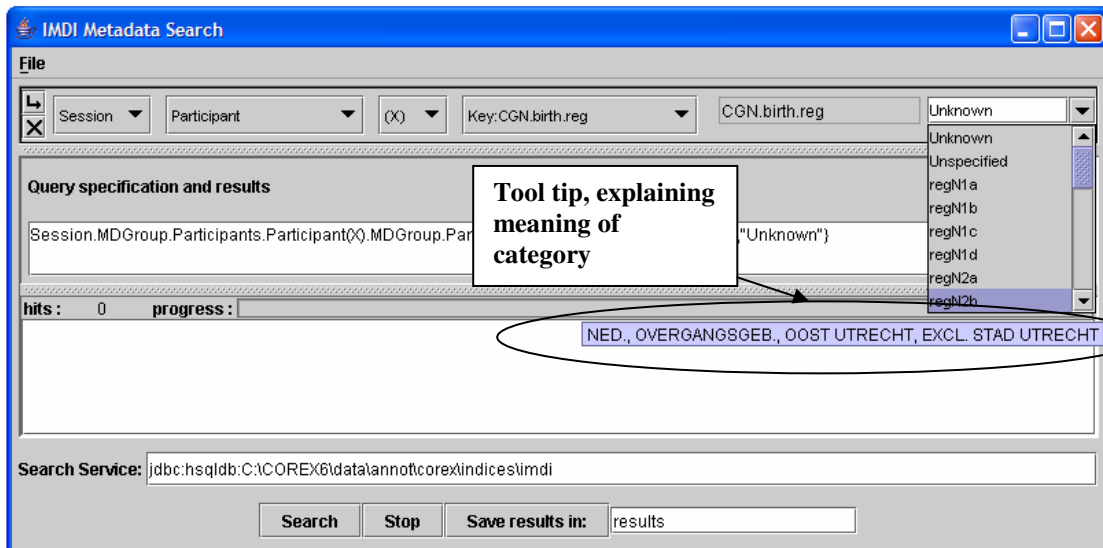
### 2.1.1 Select the category to be searched

The categories to be searched are predefined by Corex and are displayed in form of drop-down menus (see the separate CGN Documentation for an explanation of all the different categories).

The default categories, which are displayed every time you access the **Metadata Search** panel, are **Session** and **Name**. These two options allow you to search for the name of a session.

You can select other categories by clicking on the drop-down menus:

- The first category is always **Session**.
  - The second category is chosen from the following list:
    - Name, Title, Date, Location, Content, Participant, Media File and Annotation Unit. A choice for one of these will lead to a third drop-down menu.
- or:
- A number of CGN keys, leading directly to the query of your choice (i.e. they are not further subdivided in other categories).
- **! Note:** If you point with the mouse to a category that belongs to a controlled vocabulary (i.e. a pull-down menu with fixed categories), a blue box, the so-called tool tip, will appear to explain the meaning of that category, e.g.:




- For metadata fields that don't have a controlled vocabulary but a text field box in which you can specify your query, it is possible to use so-called **wild card characters** to broaden your query. The wild card characters are \* and ?. The character \* stands for 'any character zero to n times'. The ? stands for any character once.

In the example below the query in the Session → Name text box is 'fn\*'. This means all session names that begin with *fn*. This means all Dutch sessions.



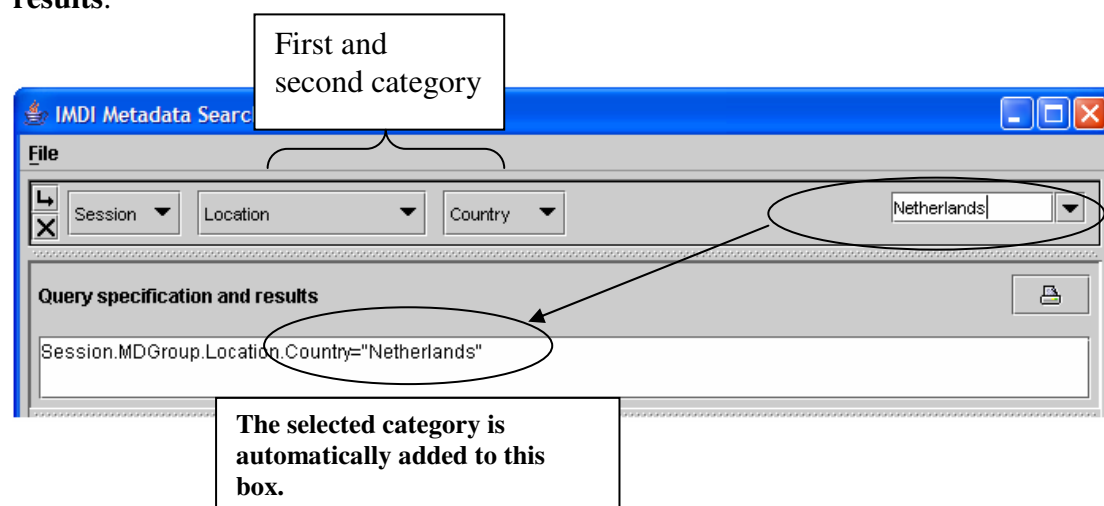
! Note: The CGN-specific keys are often mapped onto the standard IMDI data fields. However some standard IMDI routes may still lead to a dead end on the one hand, or to data fields that are irrelevant to the CGN on the other hand (e.g. searching for Location/Continent/Africa will not lead to any search results, of course). This problem can be avoided by search paths that include a CGN-specific key. The CGN-specific keys cover most of the metadata. Two main routes to access the CGN-keys are **Session/Participant/Key:** for participant information and **Session/Key:** and **Session/Content/Key:** for session information.

! Note: If too many categories and drop-down menus are added, the **Metadata Search** panel cannot display them all. To increase its size, click on the full screen icon (in the top right corner of the panel): 

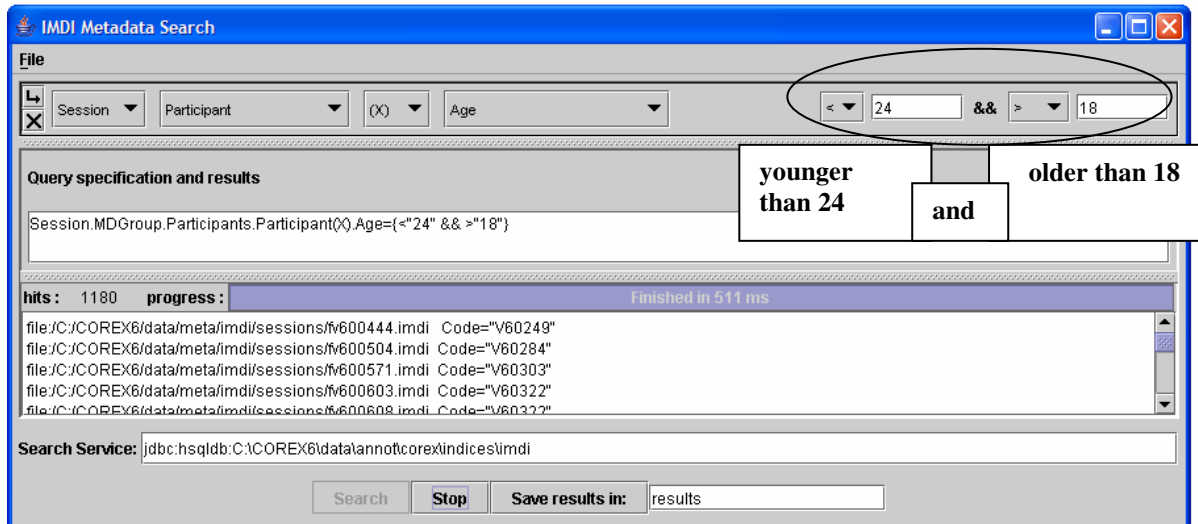
In the remainder of this section, some examples of metadata searches are given.

- (1) In the following illustration, **Location** was chosen as the second category. A third drop-down menu appears, subdividing the category **Location** into **Continent**, **Country**, **Region** and **Address**. When **Country** is chosen, a fourth drop-down menu appears, displaying the names of all countries.

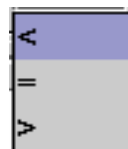
The selected category is automatically added to the box **Query specification and results**.



- (2) In another example, **Participant Age** was chosen as the second category (i.e. the age of the speaker), e.g.:



In this case, further drop-down menus allow you to specify the exact age interval with the following buttons:



search for a date before the specified date

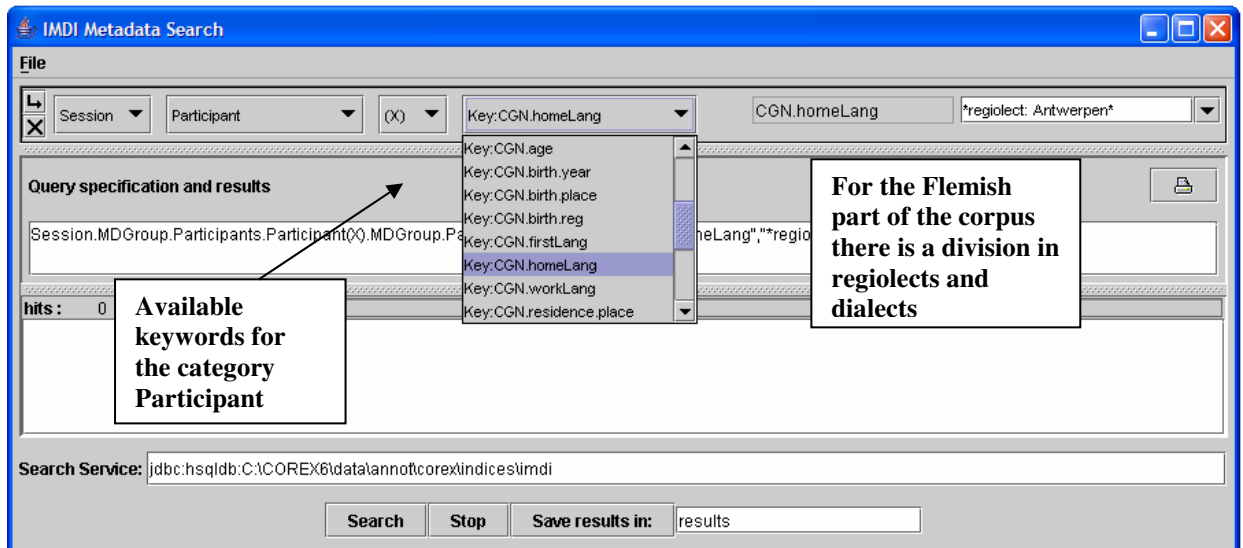
search for the exact date

search for a date after the specified date

Don't forget to press **Enter** on your keyboard after filling in a number in the text field box . Only then it becomes visible in the **Query specification and results** panel and only then it becomes effective for the query. You can also insert numbers directly in the **Query specification and results** panel. Do this only if you know what you are doing. Do not change the entire line.


- (2) The last example illustrates a search for keywords. When you select a category starting with **Key**, in most cases a drop-down menu with the available values will appear. For example, you can search in the category **Participant** for all those sessions that contain the value **SD** (means Standard Dutch) under the keyword **CGN.homeLang** (i.e. language spoken at home):





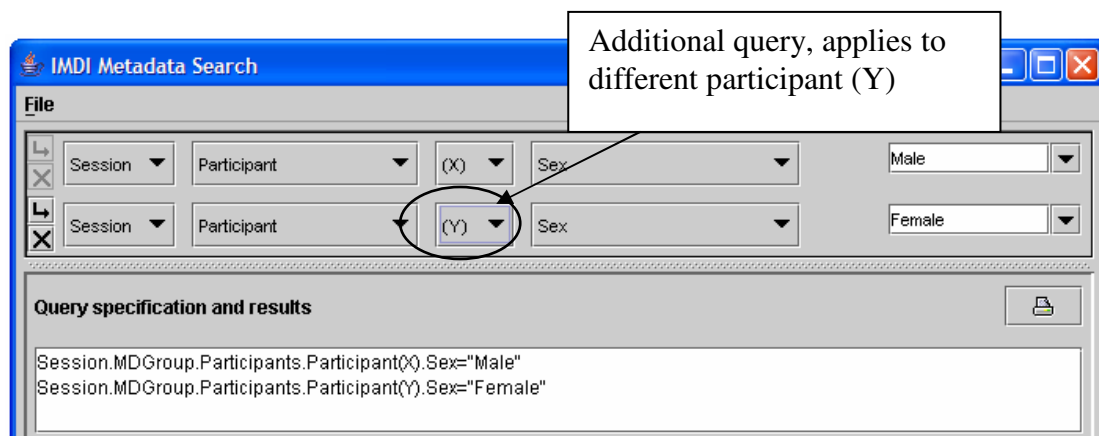
! The language keys (CGN.firstLang, CGN.homeLang, CGN.workLang) of the Flemish part of the CGN are many. They are subdivided in dialects and regiolects. For the Dutch part, the only division is between **Standard Dutch (SD)** and **Unknown**. The Flemish language keys are so many that the pull-down menu does not display all. However in the text box the regiolect or dialect of choice can be filled in manually between stars. For an overview of all the keys, see Appendix B.

Add or delete a search query You can add and delete queries as follows:

 use this button to add a query

 use this button to delete the query

These buttons are on the top left corner of the panel. When two or more queries are specified, metadata search will search for all sessions that contain query 1 *and* query 2 (*and* all subsequent queries). For example, in the following illustration, all sessions will be found that contain (1) a male participant X *and* (2) a female participant Y.



! Note: The variables X and Y are only of relevance if you specify more than one query. In this case, they allow you to specify that the criteria male and female apply to different persons X and Y (of course, in this case, the criteria could not possibly apply to one and the same person). So this combined query will search for sessions with a male and a female participant. Please ignore these variables if you specify one query only.

## 2.1 Initiate and stop the search

Click the **Search** button at the bottom of the **Metadata Search** panel to start the search. During the search the number of “hits” (matches) is shown, as well as a “bumper” moving back and forth, to indicate that the search is still in progress.

! Note: The first query may take some time, because the database has to be loaded. Subsequent queries are faster.

! Note: The name of the database that contains all metadata descriptions is specified following the coloured text box **Search Service**:. It is already configured correctly: please do not change its location or name.

Once the search process is started, you can use the **Stop** button to stop the search, for instance if you need only a few hits.

## 2.2 Display the search results

Once the search process is started, the box **Query specification and results** in the **Metadata Search** panel starts displaying the session node location for each hit.

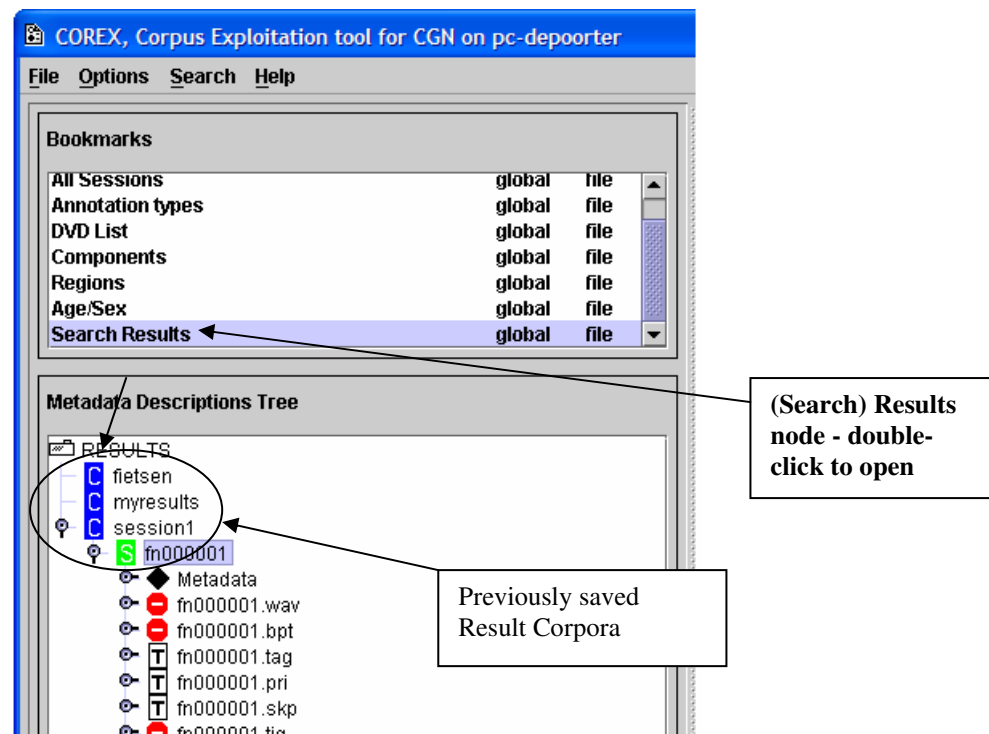
Double-click on a session node location to open a second **Corpus Browser** window that contains the selected session.

## 2.3 Save the search results

You can save all sessions that contain the search results as a corpus node. This node is displayed in the **Corpus Browser** window. Do the following:

1. In the **Metadata Search** panel, specify a name in the text box following the button **Save results in**:.:
2. Click the button **Save results in**:. The results are saved under the specified name.

You can access the results via a shortcut in the **Bookmarks** panel. Double-click on **Search Results**. The **Metadata Descriptions Tree** panel will display the new corpus under the node **Results**, e.g.:



This new corpus node “*session 1*” is treated like any other corpus node: it contains sessions (those sessions that matched the search criteria), and the sessions contain metadata descriptions and give access to all the functionality of Corex. This means that content searches, metadata searches and statistical queries can be restricted to the user-defined subcorpus.

! Note: You can do a content search on this newly created corpus (see section 4). In this case, if the newly created corpus is based on the category Participant (e.g. a corpus of all sessions that contain annotation units by male speakers below 30 years of age; see section 2.1.1), the content search on this corpus will only search the annotation units of the corresponding participant category (i.e. male speakers below 30 years of age) - if the corpus contains sessions with dialogues between, e.g. male and female speakers, the annotation units of the female speakers will be ignored.

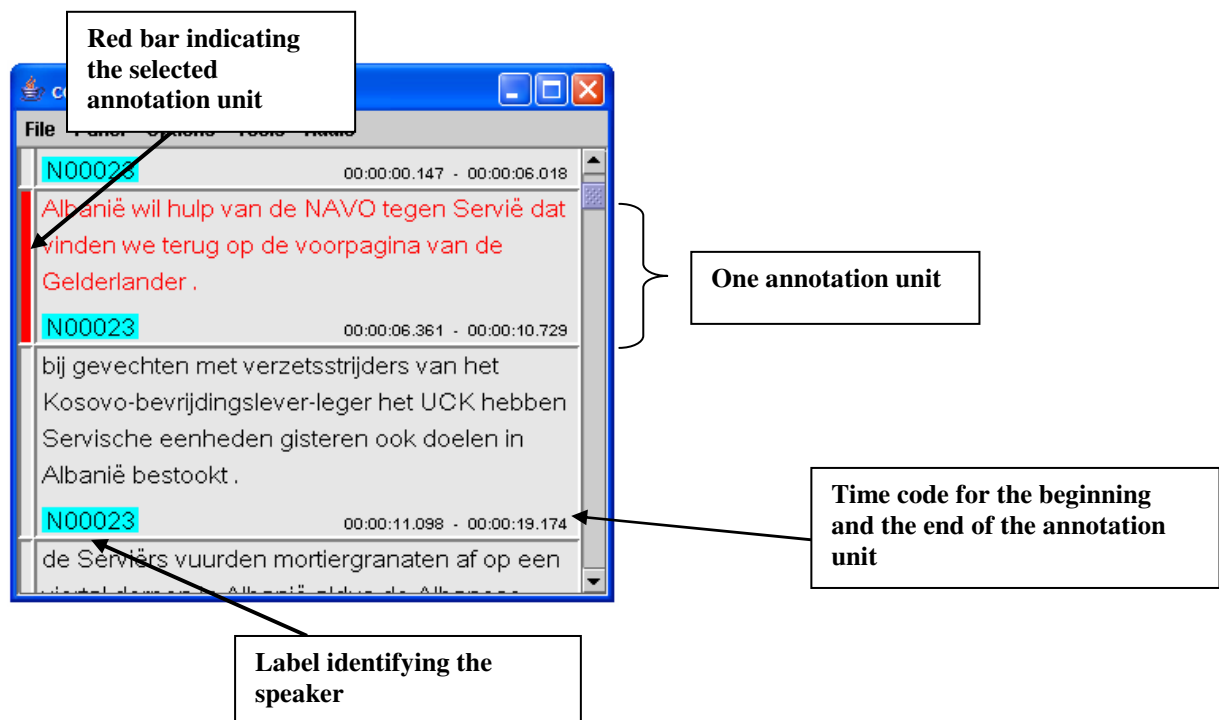
You can remove the saved corpus node again. Go to this node in the **Metadata Descriptions Tree** panel. Click on it to highlight it. Then right-click on the highlighted node and select **Remove** from the drop-down menu.

! Note: After selecting **Remove**, the icon of the removed node remains visible although it cannot be accessed anymore. The icon will be removed after exiting the node **Results**.

**! Note:** If you have too many result corpora in the search results directory, this will slow down the opening of this directory. Therefore it is important to clean up this directory regularly. The same problem is caused by result corpora that are very large.

### 3 Corex Viewer

The **Corex Viewer** is accessed by double-clicking on any session node in the **Metadata Descriptions Tree** panel of the **Corpus Browser** window (see section 1.1). It is a view panel that displays all annotation units of the corresponding session. An annotation unit (utterance) is a unit of speech bounded by a natural break in the conversation. In many cases, the annotation unit is a sentence, but in rare cases the annotation unit consists of just one word.



Use the scroll bar to the right of the **Corex Viewer** to navigate in the document.

The **Corex Viewer** is organised around annotation units. Each annotation unit contains the following information: a time code marking the beginning and the end of the annotation unit and an alphanumeric label identifying the speaker (e.g. "N00023" in the above illustration). In some cases the word **COMMENT** or **BACKGROUND** appears instead of the speaker identification number: this indicates that the annotation unit is a comment made by the transcriber.

Clicking with the mouse pointer on a word in the **Corex Viewer** selects the time segment that is associated with that word. It depends on other options how that selected segment is made visible.

A vertical red bar appearing to the left of a unit indicates it is the selected unit. Occasionally two (or more) units will have the red bars beside them (this happens when neighbouring units (partly) overlap in time and the current selected time is inside that overlap).

This section of the manual describes the options available in the **Corex Viewer**:

### 3.1 File menu

#### 3.1.1 Print view

#### 3.1.1 Print all

### 3.2 Panel menu

#### 3.2.1 Export to HTML

### 3.3 Options menu

#### 3.3.1 Show Metadata

#### 3.3.2 Play only segment

#### 3.3.3 Time Sync

#### 3.3.4 Visible Tracks

#### 3.3.5 Preferences

### 3.4 Tools menu

#### 3.4.1 Praat Synch

### 3.5 Audio menu

#### 3.5.1 Audio Player

#### 3.5.2 Waveform Panel

## **3.1 File menu**

### 3.1.1 Print view

This option allows you to print selections of text displayed in the Corex Viewer.

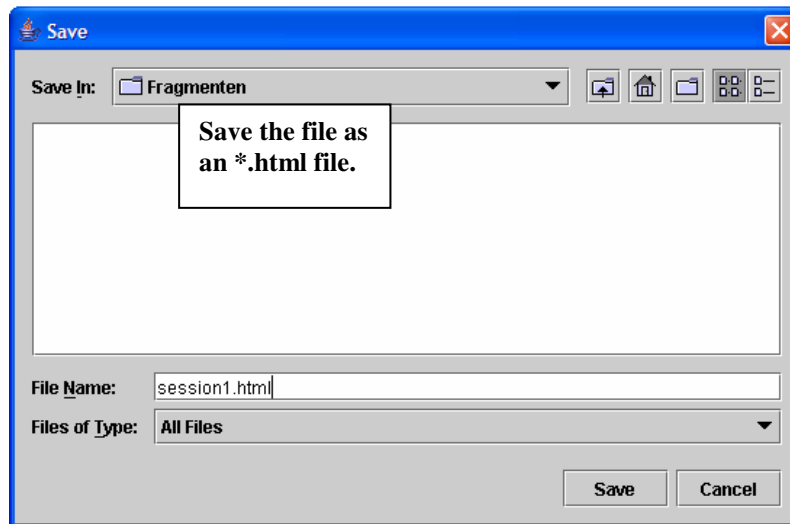
### 3.1.2 Print all

This option allows you to print the entire text.

## 3.2 Panel menu

### 3.2.1 Export to HTML

Select this option to save the content of the file. A dialogue box opens that prompts you to select the directory in which to save the file. Please save the file as an **\*.html** file, e.g.:



You can open the **\*.html** file in an HTML editor and edit it according to your wishes.

## 3.3 Options menu

### 3.3.1 Show Metadata

This option allows you to view the metadata of the opened session in the **Metadata Description Tree** panel.

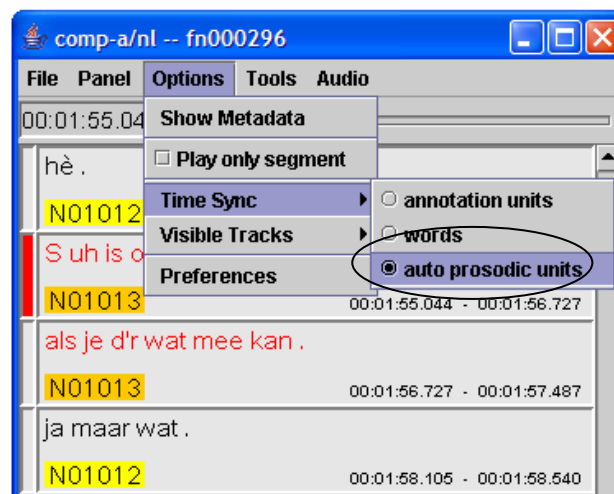
### 3.3.2 Play only segment

Restricts the audio player to the segment, i.e. the player does not play the entire audio file from the currently selected time position but only the selected segment, either a word or an annotation unit. See the next section about Time sync.

### 3.3.3 Time sync

The **Corex Viewer** can be synchronised with the **Audio Player** and the **Waveform Panel** in three ways:

- Synchronisation on the level of the annotation unit. This is the default. All the annotation units (usually just one) that contain the currently selected time are marked in red.
- Synchronisation on the level of the word. Select the option **words** under Time Sync in the **Options** menu.
- Synchronisation on the level of auto prosodic units. Select the corresponding option in the menu.

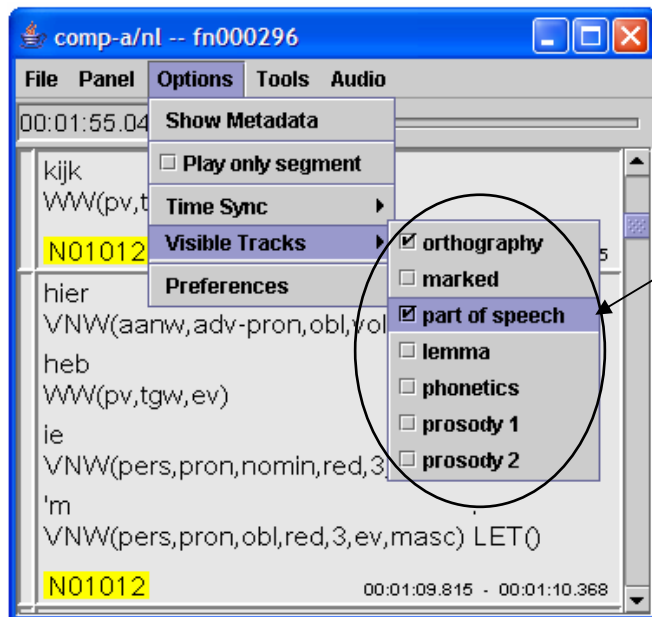


In the above example, the synchronisation was set to **auto prosodic units**. Auto prosodic units can overlap annotation unit boundaries, as can be seen in the above example. Auto prosodic units correspond to segments between automatically set (strong) prosodic boundaries.

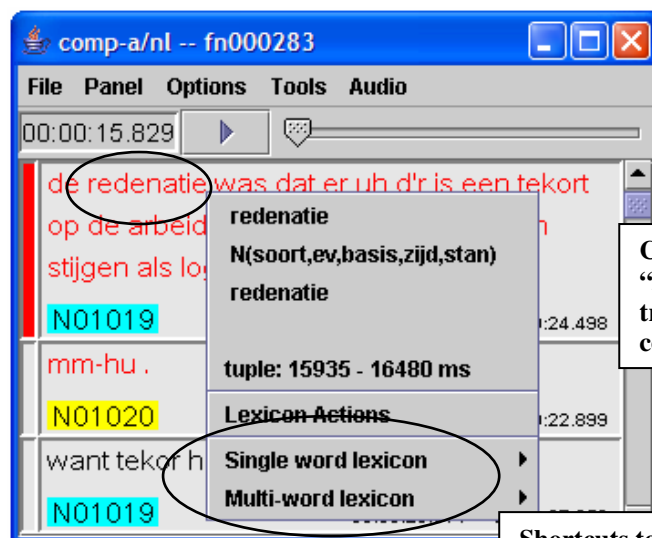
### 3.3.4 Visible tracks

Under the **Visible Tracks** option you can select the tracks (tiers) to be displayed in the **Corex Viewer**. There are up to seven different options or tracks for viewing the annotated text: orthography, part of speech, lemma, marked word, phonetics, and two types of prosody. Click on any track that you want to be displayed.





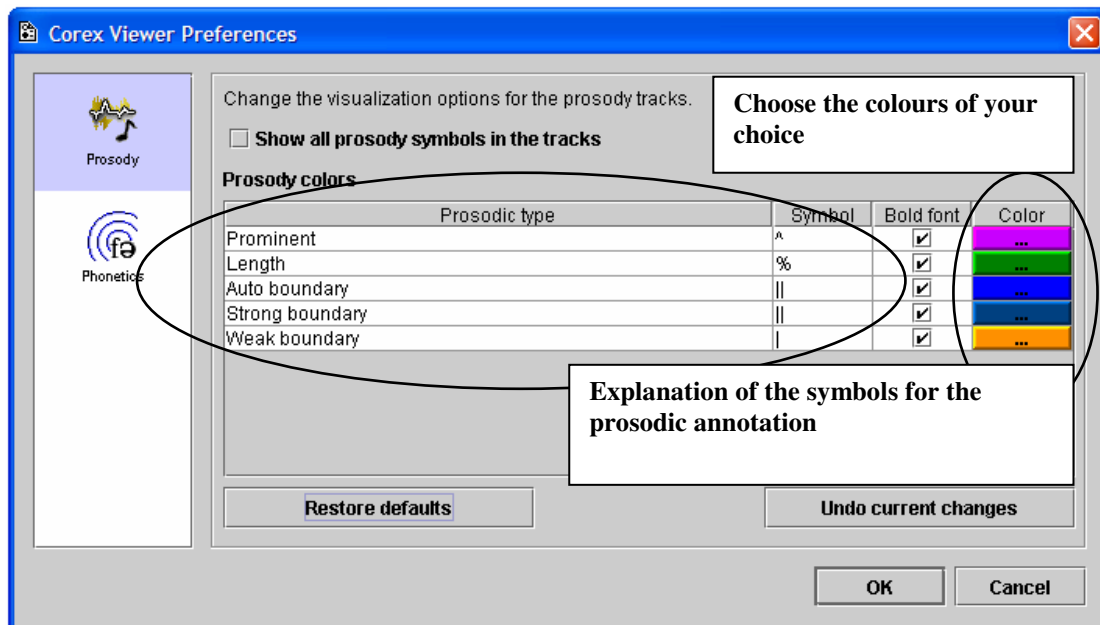
It is also possible to view all tracks (orthographic transcription, part-of-speech tag, lemma and if available phonetic transcription and/or prosodic transcription) for individual words in the annotation unit. To do so, point to the word with the mouse and click on it with the right mouse button, e.g.:



This option also allows direct lookup in the lexicon. The menu items **Single word lexicon** and **Multi-word lexicon** are shortcuts to the lexicon for either the word ID or the lemma ID of the selected word.

### 3.3.5 Preferences

This option is about viewing the meaning of the symbols for the annotation of the phonology and the prosody. Also, it provides the possibility to change the colour of the symbols for the prosodic annotation.



### 3.4 Tools menu

#### 3.4.1 Praat Synch

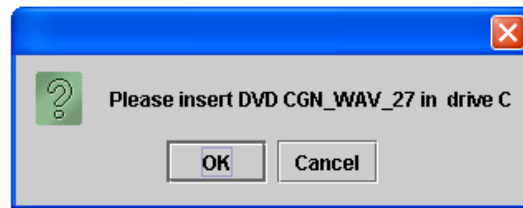
You can use the Praat signal-processing program to analyse individual annotation units. But before selecting the Praat Synch option, you must first start the Praat program, otherwise an error message appears. And you should also have selected the Audio Player option.

! Note: If you continually use the Praat signal-processing program, there will be an accumulation of data objects unless you actively remove the data objects as you finish with them. If you do not remove them, this will quickly result in a problem due to lack of available memory.

### 3.5 Audio menu

#### 3.5.1 Audio Player

When you select the Audio Player option, you are prompted to insert the corresponding DVD into the DVD drive:



At the prompt, insert the DVD and click the OK button. The **Corex Viewer** appears: there is now a new bar that contains a **media time counter** and a **scroll bar** with a play/stop button.



**Media time counter: hours:minutes:seconds.hundredths of seconds. Clicking it will turn red all text within the annotation unit segment (or word segment or prosodic segment, depending on Time sync). Also, it allows you to enter a new media time**

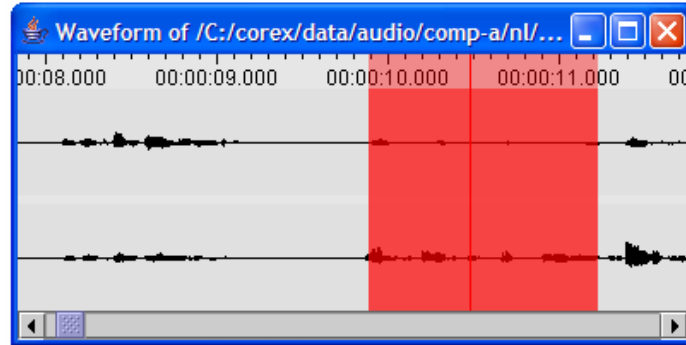
All movements of the **Audio Player** are synchronised with the **Corex Viewer** (and the **Waveform Panel**), so that any adjustments to the annotation unit location made in one will also be made in the other.

- The **media time counter** displays the media time location in the annotated text. Click on the **media time counter** box and a **Go To** window opens up allowing you to enter a new media time. Both the **Corex Viewer** and the **Audio Player** adjust their locations according to the new time you specify.
- Use the **scroll bar** to change the location.
- Click on the **play** button to begin the **Audio Player**. Once it commences, the **play** button changes to a **stop** button (a square button). As the file is played, the **Corex Viewer** displays the corresponding annotations.

**! Note:** If you are having an audio problem, first make sure that the audio volume is correctly set.

### 3.5.2 Waveform Panel

When you select the **Waveform Panel** option, a window opens up showing the corresponding waveform. The **Waveform Panel** is synchronised with the **Corex Viewer** and the **Audio Player**, so that any adjustments to the annotation unit location made in one will also be made in the other.

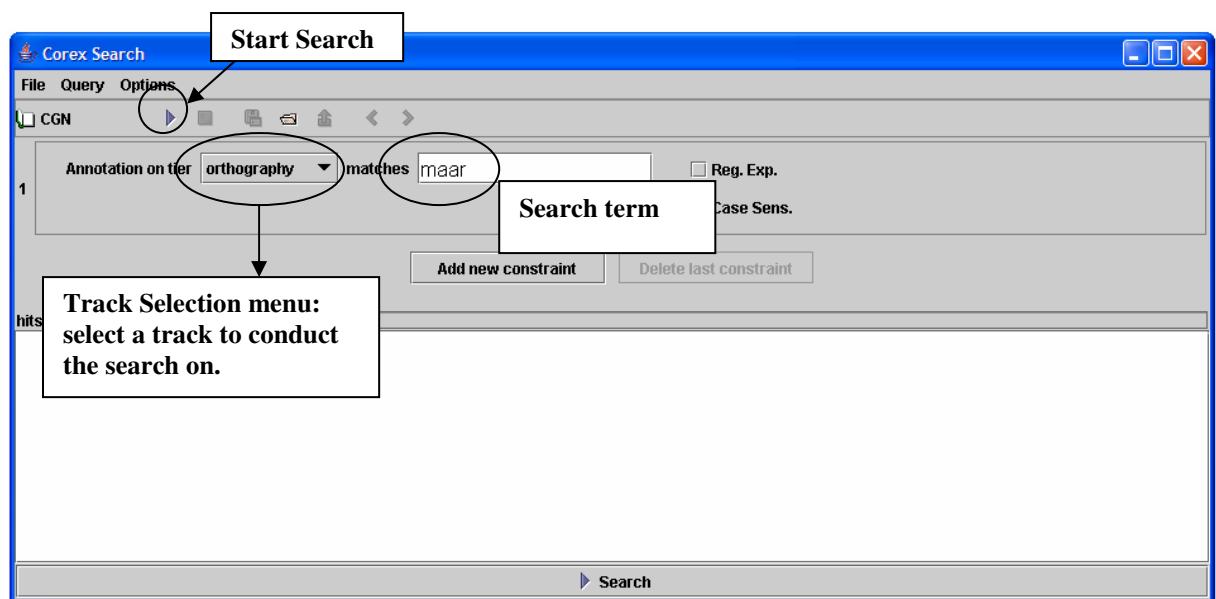


- Use the scroll bar to change the location in the waveform (and in the **Corex Viewer** and the **Audio Player**).
- A vertical red line moves along the waveform, marking the current media time. You can click on any part of the waveform to move to that location.
- You can modify the resolution of the waveform display by using a menu that appears after right-clicking in the waveform display.

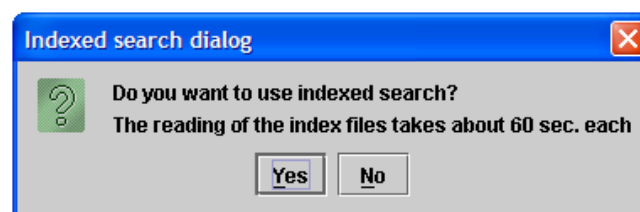
## 4 Content Search Panel

A content search is a query in the annotation files for specific tokens in a certain context. The **Content Search** panel is accessed by clicking on the pull-down menu Search and subsequently choosing Content Search.

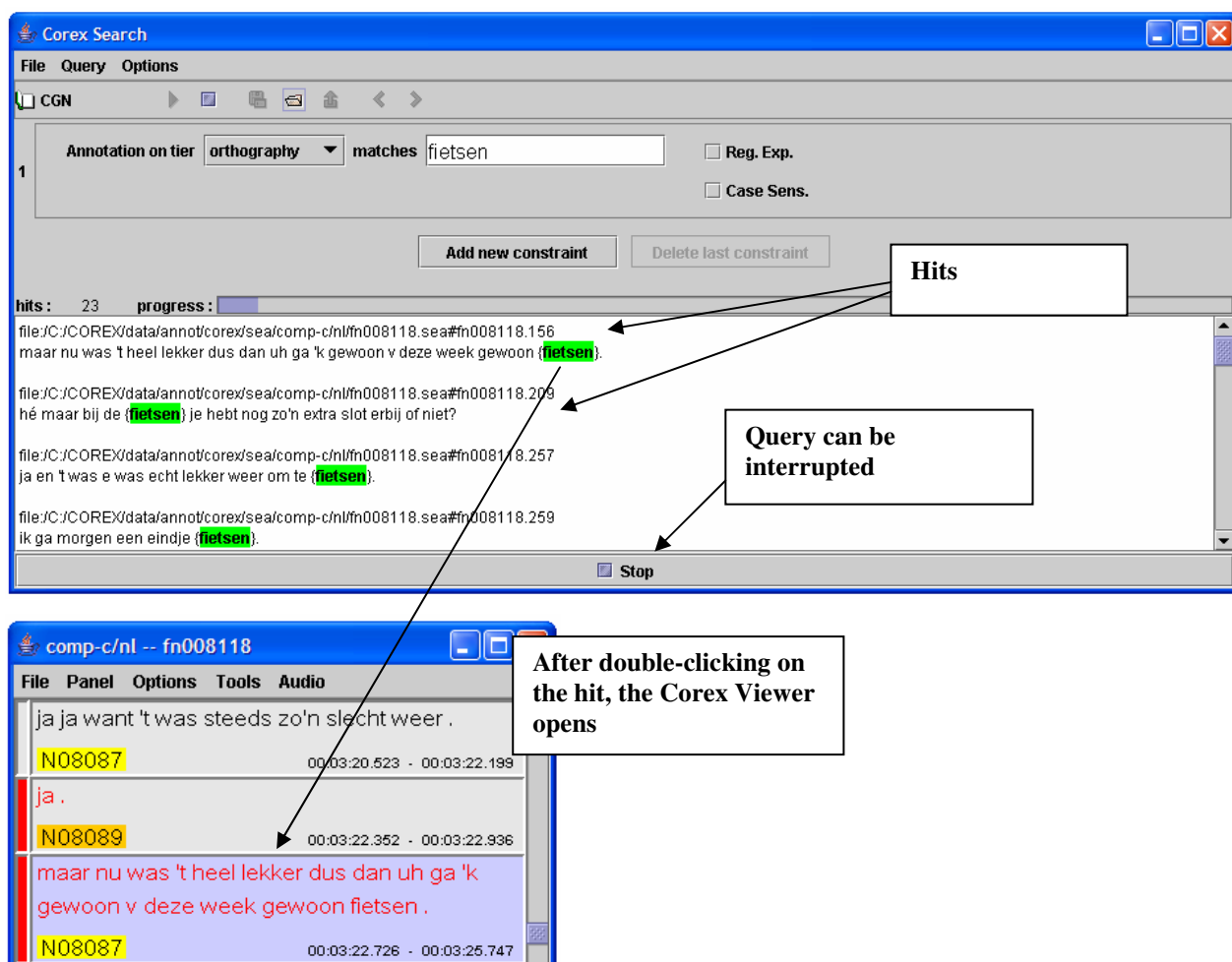
The **Content Search** panel gives you considerable flexibility for conducting searches, allowing you to specify several search parameters. Searches can be conducted throughout the entire Corpus Gesproken Nederlands (the default setting) or throughout selected parts of the corpus (see section [1.1.3](#)).



In the above case we perform a simple one-constraint search for the occurrences of the word (orthographic tier) *maar*. After clicking the Start button (or choosing from the pull-down menu Query → Search) the following panel appears:



In the dialog window, click **Yes** if you want to use an indexed search, otherwise click **No**. An indexed search is much quicker than a non-indexed search. However, the reading of the index files takes up a lot of memory. This is why – for systems that do not have a large amount of memory - the indices are deleted from memory after you pop down the **Content Search** panel. This saves memory space but has the consequence that the above panel appears each time you start a new content search.



As the query is being executed, the number of “hits” (matches) to your search criterion is shown, as well as the progress towards the completion of the search.

Once a search is completed, the session node location is given for each hit. All the hits are displayed and highlighted in green colour in the **Content Search** panel.

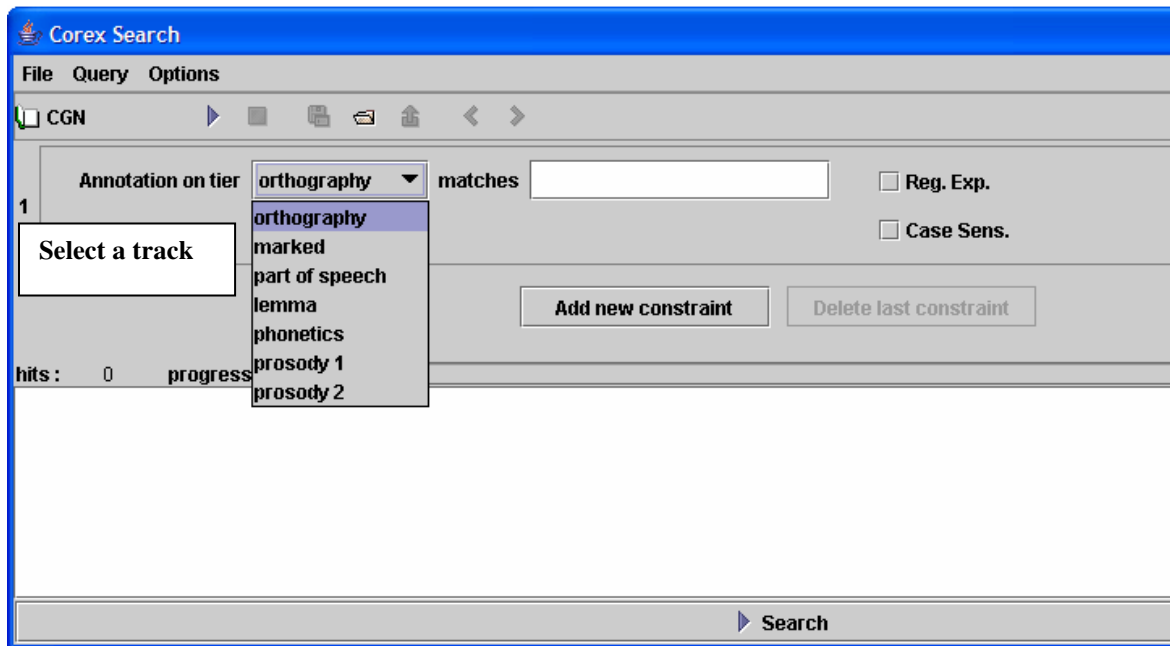
**! Note:** Because of the large amount of data that is searched, it may take some time to complete the query.

Double-click on a hit highlighted in green colour to open the corresponding **CoreX Viewer** where the annotation unit containing the search item is shown (highlighted in blue).

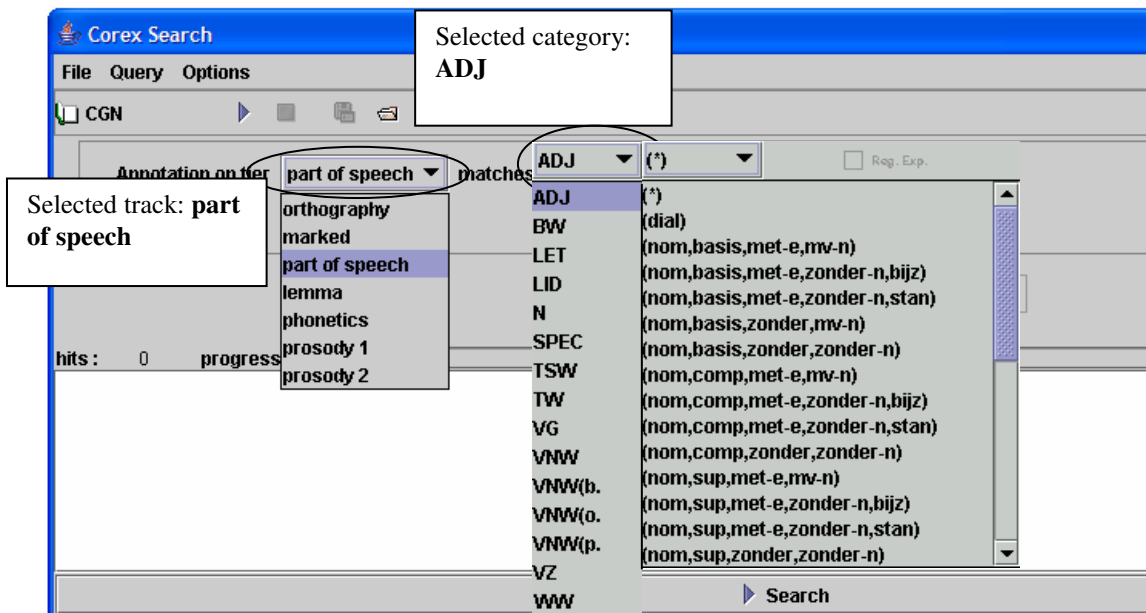
#### 4.1 Track Selection Menu and Text Field Box

Located next to **Annotation on tier** is the drop-down menu Track Selection labelled orthography (the default track or tier). To search on a track other than orthography, click on the drop-down menu and select marked (which stands for marked word),

part of speech, lemma, phonetics, prosody 1 or prosody 2. Then type in the term(s) you are looking for in the **text field box**, e.g.:



If you have selected either the track **part of speech** or the track **marked**, the **text field box** changes to a drop-down menu. Select one of the listed items (see the separate CGN documentation for details of the available categories), e.g.:



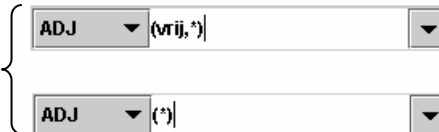
! Note: You can use an asterisk (i.e. a wild card that matches any search term) instead of one of the elements specified in the drop-down menu. Do the following:

1. Select a category, and click with the right mouse button into the drop-down menu. The drop-down menu changes into a text box.
2. Click with the left mouse button into the text box.
3. Press the key DELETE. The last element is replaced with an asterisk. Only the last elements can be deleted, not middle or first elements.
4. Repeat steps (2) and (3) to replace other elements.

Step 1: 

Step 2: 

Step 3: 

Step 4: 

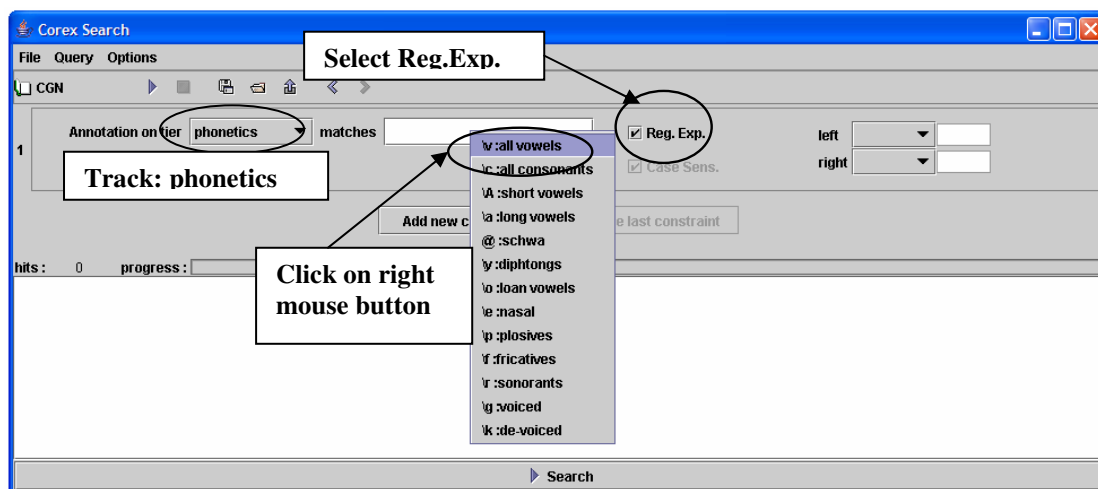
#### 4.1.1 Phonetics

A part of the corpus has a verified phonetic transcription.

The phonetically verified part can be exclusively accessed by choosing the **Annotation types** → **phonetic annotations** option in the **Bookmarks** panel.

- Queries can be conducted within this data. If you select the track **phonetics**, you can type in phonetic characters. You can use the CGN symbol set (derived from the SAMPA (Speech Assessment Methods Phonetic Alphabet) set.) For example, the SAMPA symbol @ matches the phoneme *schwa*. A search for “*h@r*” on the track **phonetics** would thus find all instances of “*het*”.
- You can also use the predefined selection of phonemes, as is demonstrated below. To do this, click on **Reg. Exp.** and click on the right mouse button when standing above the text field box.



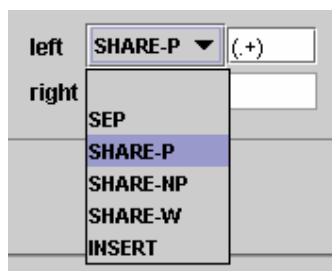


- The phonetic transcription also incorporates information about assimilations across word boundaries, degemination and insertions of phonemes between words. The following line from a **.bpt** file is an example of degemination: two phonemes are pronounced as one:

```
<fw ref="fn000021.7.2" w="en" fon="En=" left="SEP" right="SHARE-NP(n)" fq="man" times=""/>
<fw ref="fn000021.7.3" w="nieuwe" fon="=niw@" left="SHARE-NP(n)" right="SEP" fq="man" times=""/>
```

Look at the values of the data fields *left* and *right*, which are of our concern now. The words ‘en’ and ‘nieuwe’ share the phoneme ‘n’. This is a non-plosive phoneme, hence SHARE-NP. The pull-down menus on the top left of the **content search** panel - if you select the tier **phonetics** – provide the possibility to select the following options:

- SEP: the words do not share a phoneme
- SHARE-P: the words share a plosive
- SHARE-NP: the words share a non-plosive
- SHARE-W: the word “da’s” is the only instance of this category
- INSERT: inserted phonemes that are not part of any of the neighbouring words
- If you don’t select any of the above categories, the search does not care about any boundary condition and includes all values it can take.



In addition to the general boundary category, a specific boundary phoneme can also be included in the search. In the above example this is the phoneme *n*. Combined with the SHARE-NP condition for the left boundary this will result in hits such as ‘in negen gevallen’ where the phoneme *n* of *negen* is shared with the previous word.

A regular expression is allowed in the text field box. For example, the expression ([nm]) will search for either an ‘m’ or an ‘n’. By default, the text box has the expression (.+) in it. This simply means that if there are no restrictions, it matches any character or string of characters (See section [4.2](#)).

#### 4.1.2 Prosodic search

A part of the corpus is annotated for prosody.

- The prosodic part of the CGN can be exclusively accessed by choosing the **Annotation types** → **prosodic annotations** option in the **Bookmarks** panel.
- Prosodic units can extend beyond the boundary between two annotation units. It is possible in Corex to apply viewing options and search options to the prosodic unit, in addition to the word unit and the annotation unit (see section [3.3.3](#) and section [4.4](#)).
- The annotation was performed according to the four different standards of annotating prosody that exist in the Netherlands and Flanders (two for each country). This means that for each annotation unit, which is either Dutch or Flemish, a choice between two options must be made: prosody1 or prosody2.
- For a colour coding of the prosodic symbols, choose **Options** → **Preferences** in the Corex viewer. The below picture is an extraction of the panel that will become visible.

Prosodic type	Symbol	Bold font	Color
Prominent	^	<input checked="" type="checkbox"/>	...
Length	%	<input checked="" type="checkbox"/>	...
Auto boundary		<input checked="" type="checkbox"/>	...
Strong boundary		<input checked="" type="checkbox"/>	...
Weak boundary		<input checked="" type="checkbox"/>	...

- The boundaries are classified as either weak or strong boundaries. Weak boundaries are by definition the result of a manual transcription, whereas strong boundaries are either the result of a manual transcription (**Strong boundary**) or the result of the automatic transcription (**Auto boundary**).
- The prominence of a syllable is denoted by the symbol ^ in the **.prx** files and with a purple colour in the **Corex Viewer** (in the default settings - these can be changed).
- The extended length of a syllable is denoted by the % sign in the **.prx** files and by a green colour (default) in the **Corex Viewer**.



In the above example of a Dutch fragment, the tier Prosody1 is visible.

An example of a line from a **.prx** file is the following:

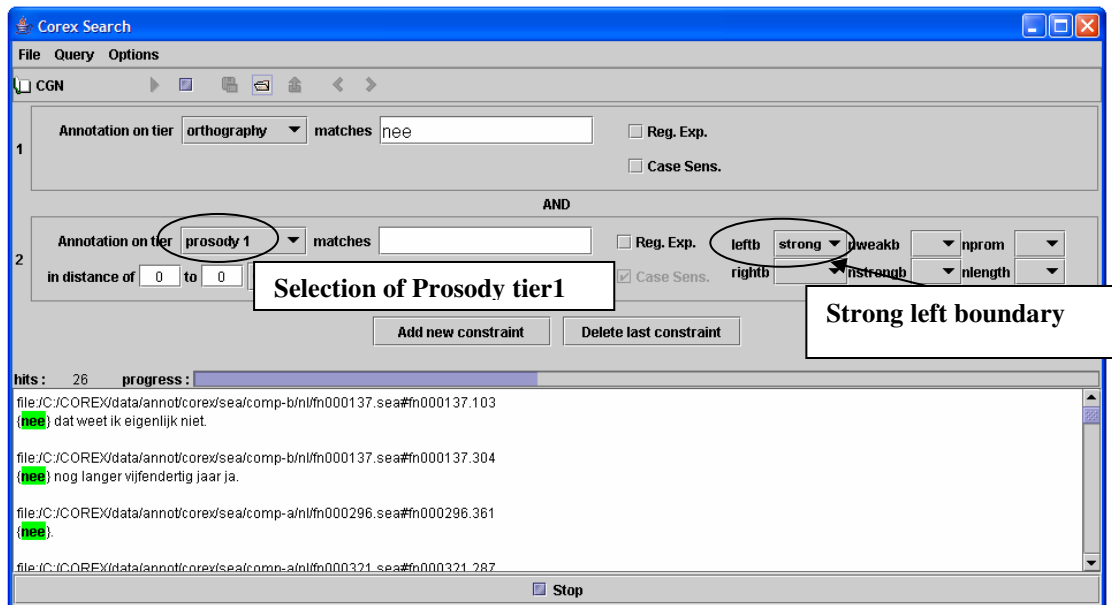
```
<prw ref="fn000055.1.5" w="bijdrage" annot="b^ij^drage" nprom="1" nlength="0" nweakb="0" nstrongb="0"
```

It shows that the word 'bijdrage' has one prominent unit (the value of nprom is 1) and no weak or strong boundaries and no extended length. This way of coding the prosody gives us the opportunity to perform queries in the prosodic part of the CGN.



The options left boundary and right boundary can be set to different values, as is shown above. Also, the number of weak and/or strong boundaries can be chosen. The same goes for the prominence or length.

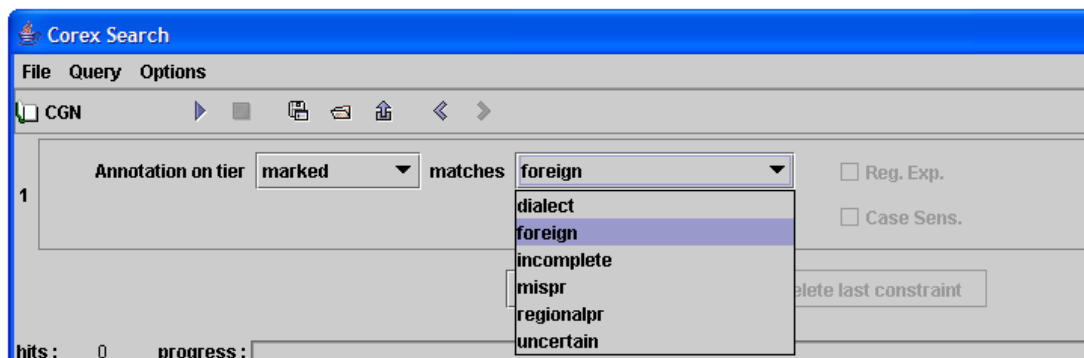
The values of these variables can for example be used as an additional constraint on another condition (see below).



In the above example, a prosodic search was conducted. More in particular, a query was done for all occurrences of *nee* (first constraint) preceded by a strong left boundary (second constraint).

#### 4.1.3 Marked Words

Some words are marked (in the **.pri** file). The categories are: dialect, foreign, incomplete, mispronounced, regional pronunciation and uncertain. See the protocol for orthographic transcription for a detailed explanation.



## 4.2 Regular Expression Search

If you type in a letter string in the **text field box**, this will result in matches in the orthographic tier for *exactly that letter string*. However, it is also possible to refine your search term in such a way that it captures *classes of letter strings*. This is called regular expression search and it is an option in all text field boxes of Content Search.

A regular expression search makes use of variables, using Perl syntax, e.g.:

- . matches an arbitrary character
- + matches the preceding pattern element one or more times
- ? matches the previous character zero or one times
- \* matches the previous character zero or more times
- ^ matches the beginning of a word
- \$ matches the end of a word
- \b matches the beginning or end of a letter string
- [adg] matches the character a, d or g
- [^ad] matches any character except a and d

A more elaborate survey of this regular expression syntax can be found in numerous books and websites<sup>1</sup> on the programming language Perl.

To enable a regular expression search, do the following:

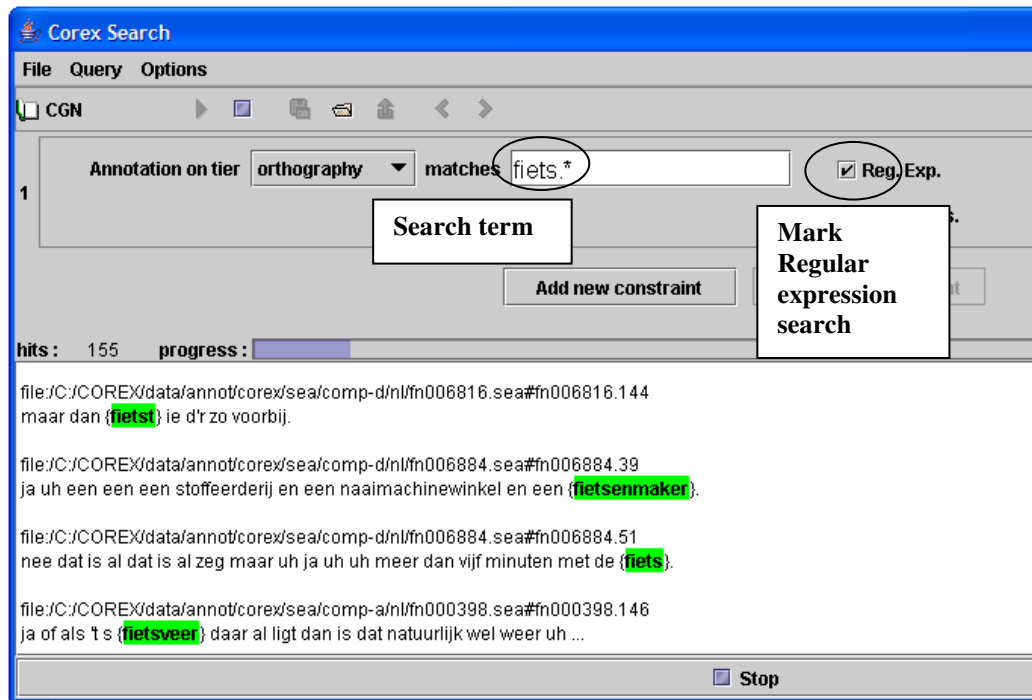
1. Enter the search term(s) in the **text field box**
2. Click in the box to the left of **Reg. Exp.**

A checkmark appears in the box indicating that this option has been selected.

The following example illustrates a regular expression search for “*fiets.\**”. Note that the search results not only include “*fiets*”, but also, e.g., “*fietsenmaker*” and “*fiets!*”.

---

<sup>1</sup> <http://www.english.uga.edu/humcomp/perl/regex2a.html>  
<http://www.wellho.net/regex/perl.html>



### 4.3 Case-sensitive search

By default, any search is case-insensitive. If you want to conduct a case-sensitive search, do the following:

1. Enter the search term(s) in the **text field box** (paying attention to upper and lower case letters).
2. Click in the box to the left of **Case Sens.** A checkmark appears in the box indicating that this option has been selected.

E.g., it is possible to search for just those instances of the search term “*Het*” that start with an upper case letter.

### 4.4 Add new constraint button

The Add new constraint button allows you to modify the initial query by adding other conditions. You can add as many conditions as necessary.

To add search parameters, do the following:

1. Specify the first search condition (as described in sections [4.1](#) to [4.3](#)).
2. Click on the button Add new constraint. A new constraint is added. The new constraint is always combined to the former with an AND-operator.
3. Specify the second search condition (as described in sections [4.1](#) to [4.3](#)).

4. Specify the relation between the first and the second condition, i.e. specify how many words, annotation units or prosodic units should come between the first and the second condition.

For example: We want to select only those instances of “*zijn*” which belong to the category of possessive pronouns (hereby we exclude instances of *zijn* that belong to the category of verbs).

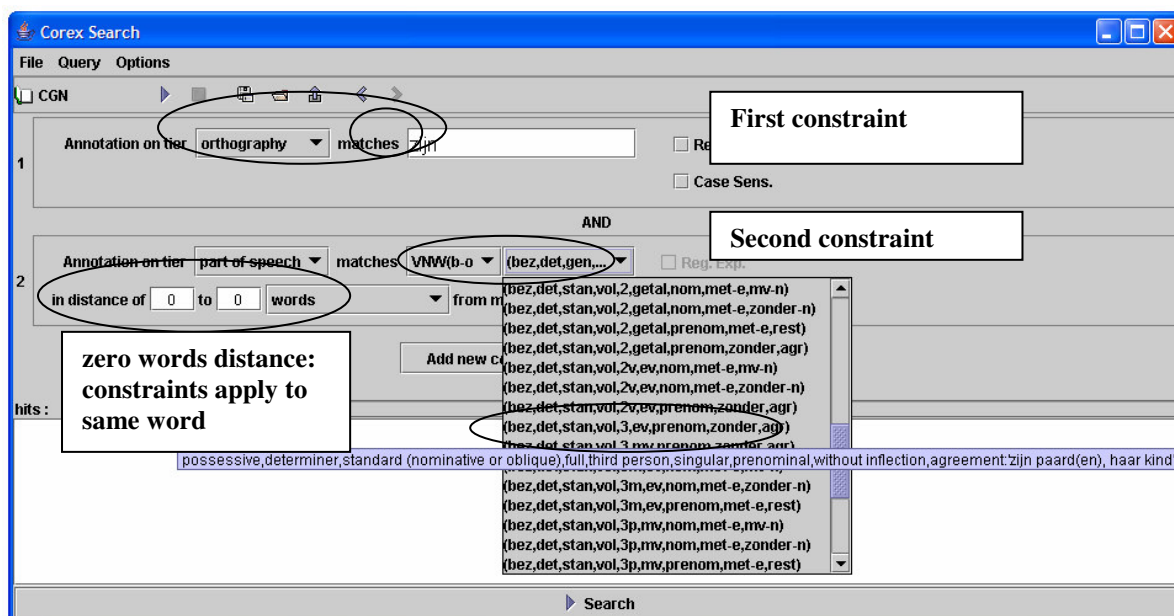
If we specify “*in distance of 0 to 0 words from matches found in 1*” (as in the screenshot below), Corex will find only those **words** that match both the first and the second parameter. In the current case these are all words “*zijn*” that belong to the part of speech “*VNW (bez, det,...)*”.

It is also possible to apply the combined constraints on words within the same *annotation unit*. This is less useful in our example, since we are looking for one word – not one annotation unit - to which multiple constraints apply. If we would specify “*in distance of 0 to 0 annotation units from annotation found in 1*”, Corex would find all those **annotation units** that match both the first and the second parameter.

A third option is to search within a range of **prosodic units** from the hit from the previous condition. Note that in this definition of prosodic units, the strong boundaries *within* words are not counted as boundaries.

If you want to add further constraints, repeat steps (2) to (4) above.

The following screenshot illustrates a search for the word “zijn” (first search condition) of the part of speech “VNW (bez, det, ...)” (second search condition). Corex finds only those words that match both conditions.

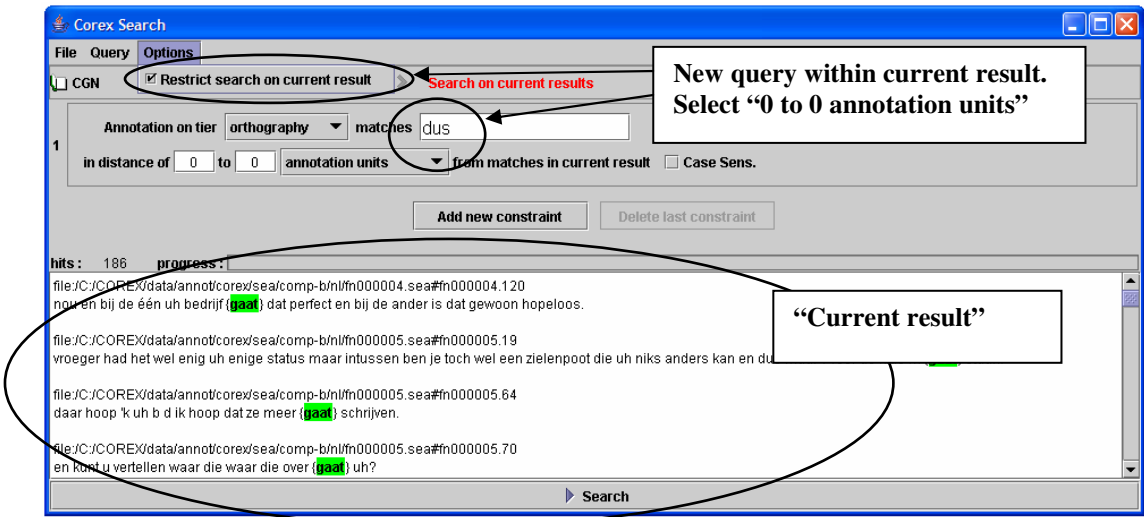


#### 4.4.1 Performing a query within results

An alternative way of combining constraints is to select the option **Restrict search on current result**. The advantage of this iterative way of using constraints is that it is faster. The disadvantage is that the overview on the constraints is less clear (you have to remember it, i.e. you cannot see which constraint applied to the prior selection). The procedure goes as follows.

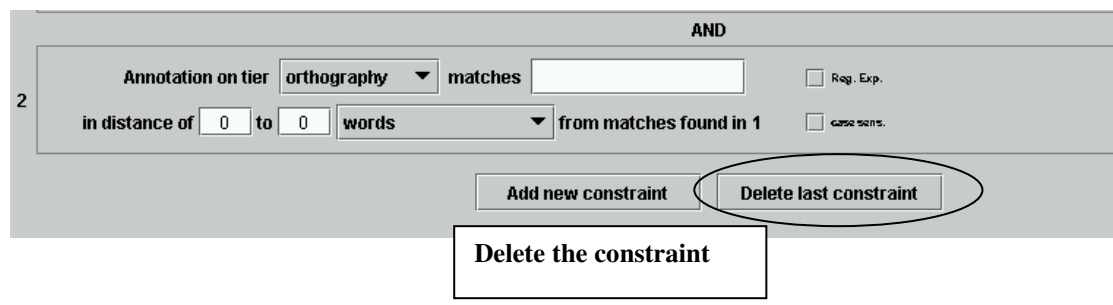
1. Perform a single constraint query. In the below example this is the search for all instances of *gaat* on the tier orthography (this query was interrupted, but that is not relevant to the example).
2. Choose the option **Restrict search on current result**.
3. Choose the second constraint. The search that is subsequently performed will apply only to the results that have been found in the first query (or a saved query, see 5.5). Don't forget to select the option **in distance of 0 to 0 annotation units**.






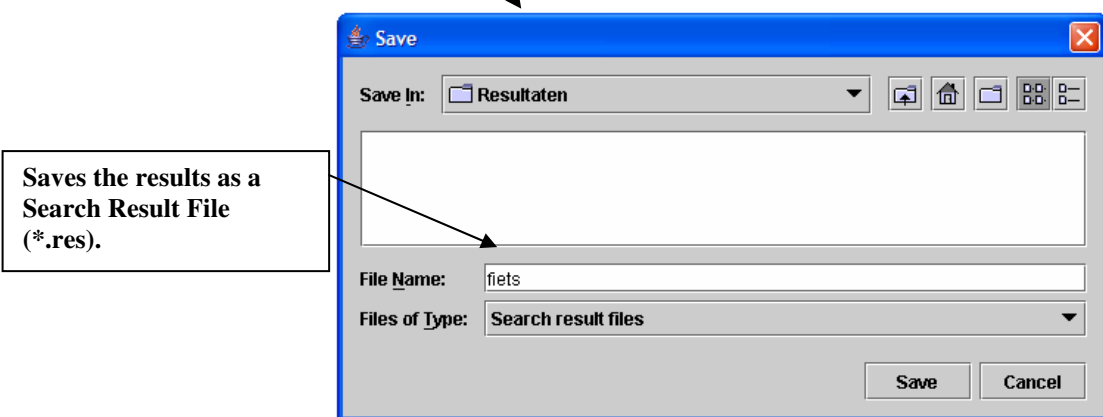
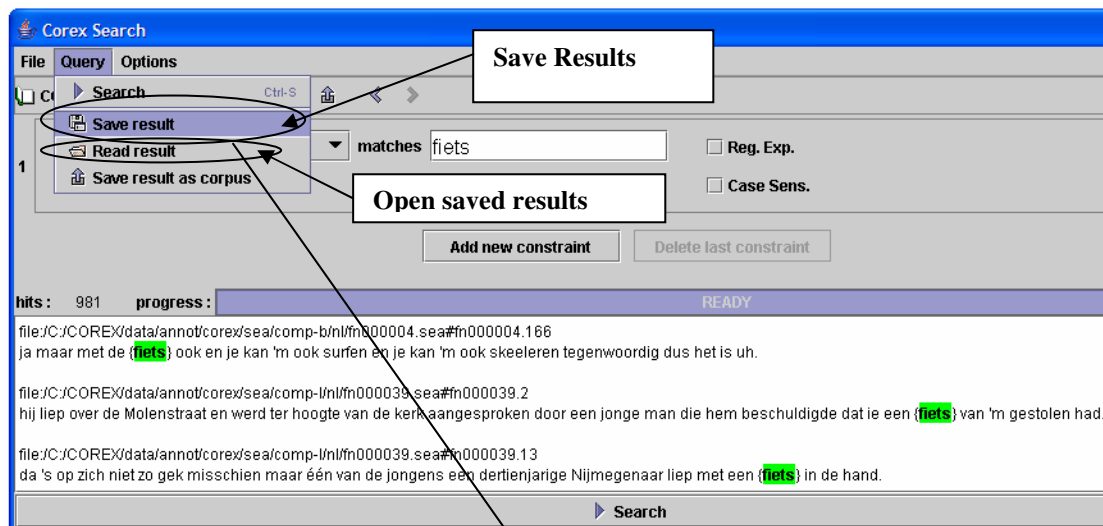
### 4.5 Delete last constraint button

Click this button to remove the last constraint. It is deleted without further warning.




### 4.6 Save Result

Click the Save Result option in the Query menu to save the results of your search. A dialogue box will open that prompts you to select the directory in which to save the file (see below). Please save it as a **Search Result File (\*.res)**. There is also a shortcut icon to this option: 



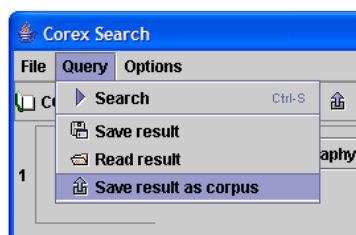
The result file contains all possible information on the hit, except the links to the metadata. The format of this file is illustrated in Appendix A. This file can, for instance, be used as a basis to extract all audio segments corresponding to the query results from the audio files.

## 4.7 Read result


Choose the Read result option in the Query menu to open a file that was saved with the Save result option (i.e. saved as a **Search Result File, \*.res**). A dialogue box will open that prompts you to browse to the desired file. There is also a shortcut icon to this option: 

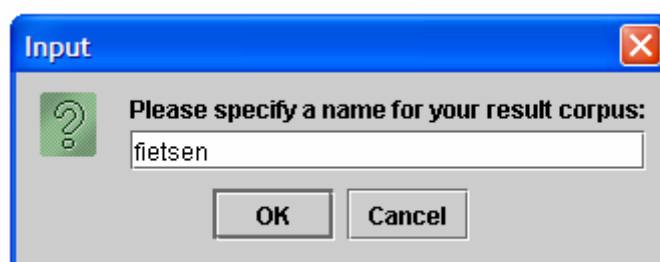
**! Note:** Any opened **\*.res** file will be displayed exactly as a normal search result, i.e. the locations of the hits are given, the hits are displayed in green colour and double-clicking on any hit will automatically open up the corresponding **Corex Viewer**.

## 4.8 Save results as corpus

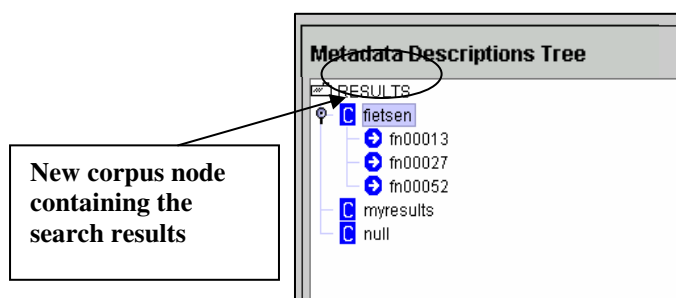


The option **Save result as corpus** is suited if a subsequent **Metadata Search** is to be performed. This is because saving the results as corpus keeps the metadata linked to the annotation data, a link that is lost if you choose the option **Save result**.

Click the **Save result as corpus** option from the **Query** menu (or use the shortcut icon ) to save all sessions that contain the search results as a corpus node in the **Corpus Browser** window. You are then asked to specify a name for the new corpus, e.g.:



Click **OK** to save the specified corpus under the node **Search Results** in the **Corpus Browser** window (see section [1.2](#)). The **Metadata Descriptions Tree** panel will display the new corpus node, e.g.:



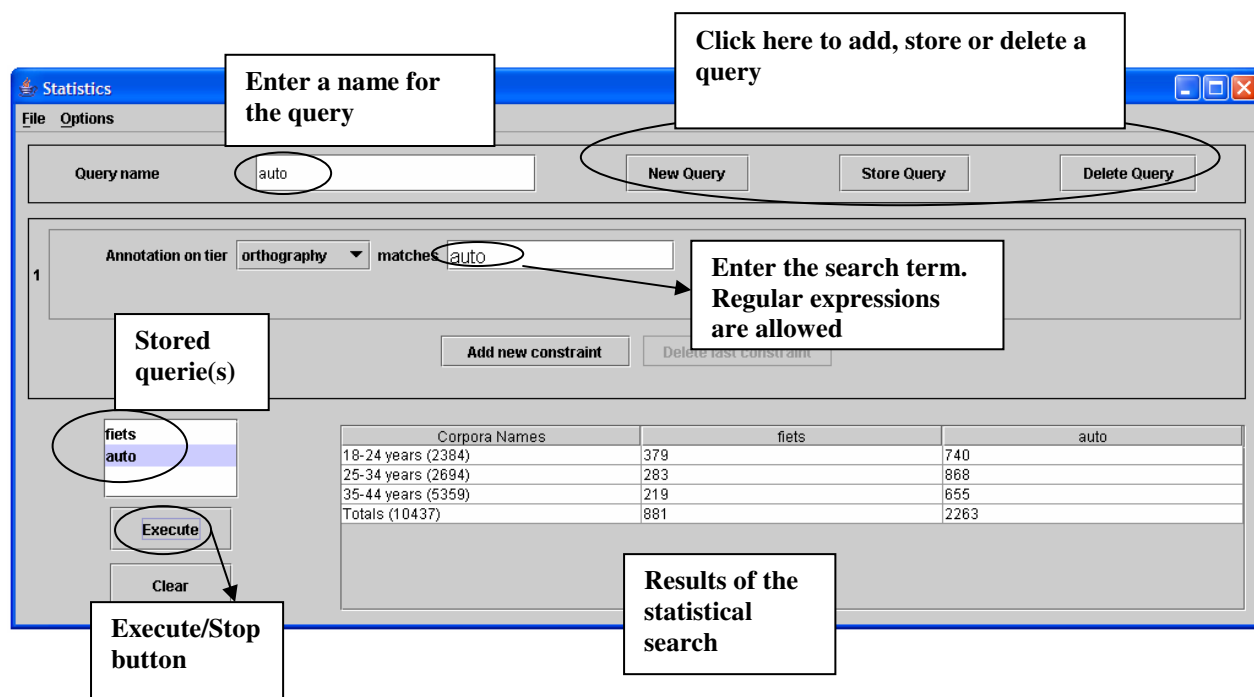
This new corpus node is treated like any other result corpus node: it contains sessions (those sessions where the search term was found), and the sessions contain metadata descriptions and give access to the **Corex Viewer**, the **TIGERGraph Viewer**, the **Audio Player** and the **Waveform Panel**.

You can remove the saved corpus node again. Go to the node in the **Metadata Descriptions Tree** panel. Click on it to highlight it. Then right-click on the highlighted node and select **Remove** from the drop-down menu.

! Note: After selecting **Remove**, the icon of the removed node remains visible although it cannot be accessed anymore. The icon will be removed after exiting the node **Results**.

## 5 Statistics Panel

The **Statistics** panel allows you to count occurrences of a term throughout the Corpus Gesproken Nederlands (CGN). To access the **Statistics** panel, click on the button menu option **Statistics** in the **Search** menu. The **Statistics** panel looks as follows:



- An important feature of performing a statistical query is that multiple selections can be added to the basket and that the counts for these selections are displayed separately (see the above example).

To conduct a statistical search, do the following:

1. In the **Metadata Descriptions Tree** panel, select the corpus or corpora that you want to search (see section [1.1.3](#)). You have to select a corpus; there is no default corpus in the basket.
2. In the **Statistics** panel, enter a name for your query in the box **Query name**. Specify your search options and enter the search term in the **text field box** (see section [4.2](#) for details on how to specify your search options).
3. Click on the button **Store query**. The query will be stored and displayed in the box at the bottom left side of the **Statistics** panel.

If you want to add other queries, click on the button **New query**, and repeat steps (2) to (4) above.

If you want to delete a query, click on the corresponding query displayed in the box at the bottom left side of the **Statistics** panel. Then click on the button **Delete query**. The query is deleted without further warning.

If you want to delete all queries, click on the button **Clear**. All queries are deleted without further warning.

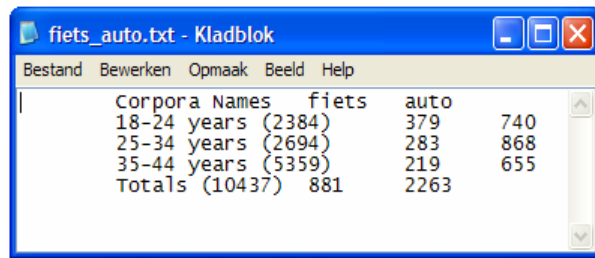
4. Click on the button **Execute** to start the search.

The selected corpus or corpora is/are searched for the number of occurrences that match your query. During the search, the **Execute** button switches to a **Stop** button that allows you to stop the search in progress.

The results of the search are displayed in a table format, separately for each query and each corpus. E.g. in the screenshot above, the occurrences of “*fiets*” and “*auto*” are counted separately for each specified corpus (i.e., for the corpus “*18-24 years (containing 1737 sessions)*”, for the corpus “*25-34 years (containing 1948 sessions)*” and for the corpus “*35-44 years (containing 4716 sessions)*”).

5. Once the search is completed, you have the following options:
  - Click on the menu item **Show Totals** in the **Options** menu to view the total number of matches for each query.
  - Click on the menu item **Save Results** in the **File** menu to save the results.

The saved file can be opened in any program that can handle tab-delimited texts, e.g. in Microsoft Excel.



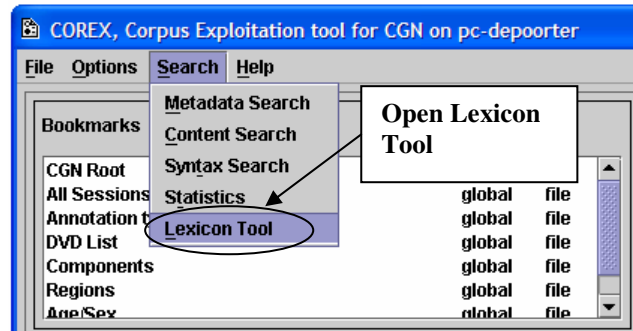
The screenshot shows a Notepad window titled 'fiets\_auto.txt - Kladblok'. The window contains a table with the following data:

Corpora Names	fiets	auto
18-24 years (2384)	379	740
25-34 years (2694)	283	868
35-44 years (5359)	219	655
Totals (10437)	881	2263

6. When you are finished with the statistics search, click on the **Exit** menu option in the **File** menu to close the **Statistics** panel and return to the **Corpus Browser** window.

## 6 The Lexicon Tool Panel

The lexicon tool is accessed by choosing the Lexicon tool option from the Search menu.



The Lexicon is a source of additional lexical information. A result of a query in the lexicon can serve as a starting point for a subsequent query in the CGN annotations.

- All orthographic entries of the CGN (and hence lemma entries and part-of-speech entries) are incorporated in the lexicon, and the lexicon contains only these.
- The CGN entries are supplied with idealized lexical information, such as standard pronunciation.
- Multi-word expressions are also available: the Multi-word Lexicon.

### 6.1 Performing a Query in the Single Word Lexicon

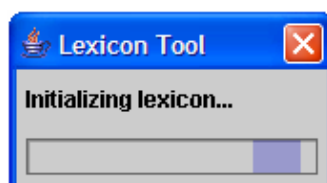
In the below example, a query is performed on the orthographic field. Click on the field selection menu to choose a different field.

The screenshot shows the Lexicon Tool interface with the following components:

- Search Results:** A table listing search results for the term 'aan.\*' in the Orthography field. The table has columns for Orthography, Part of s..., Lemma, Sy..., U..., Pr..., Pr..., Pr..., P..., M..., D..., and C... The results include words like 'aan', 'aan-uitschakel...', 'aanbel', 'aanbelangt', 'aanbelde', 'aanbelden', 'aanbellen', 'aanbelt', 'aanbesteder', 'aanbesteding', 'aanbestedingen', 'aanbestedings...', 'aanbestede', and 'aanbетаald'.
- Lexicon type:** A section with two radio buttons: 'Single word lexicon' (selected) and 'Multi-word lexicon'.
- Search term:** A text box containing 'aan.\*' and a 'Find' button.
- Field selection menu:** A dropdown menu showing 'Orthography' selected.
- Buttons:** 'Search in Corpus', 'Statistics in Corpus', 'Export results', and 'Print results'.
- Other elements:** 'Nr of hits: 1918', 'within results' checkbox, and a 'Field selection menu' label pointing to the dropdown.

- Choose **Single word lexicon** as the lexicon type (this can also be done by choosing File → Lexicon Type → Single word lexicon.)
- Choose the field of your choice. In the current example this is **orthography**.
- Enter the search term in the **text field box**. It is possible, of course, to conduct a search without using a regular expression. However, in the present example the search is a regular expression. 'aan.\*' is a query for all words that start with 'aan' (the syntax of regular expressions is Perl syntax, as demonstrated in section 4.2 about content search. However, contrary to the content search procedure, no check box Reg. Exp. needs to be clicked here).
- Click on the button **Find** to start the query.

The first time of each session the lexicon is initialised and the below panel appears.

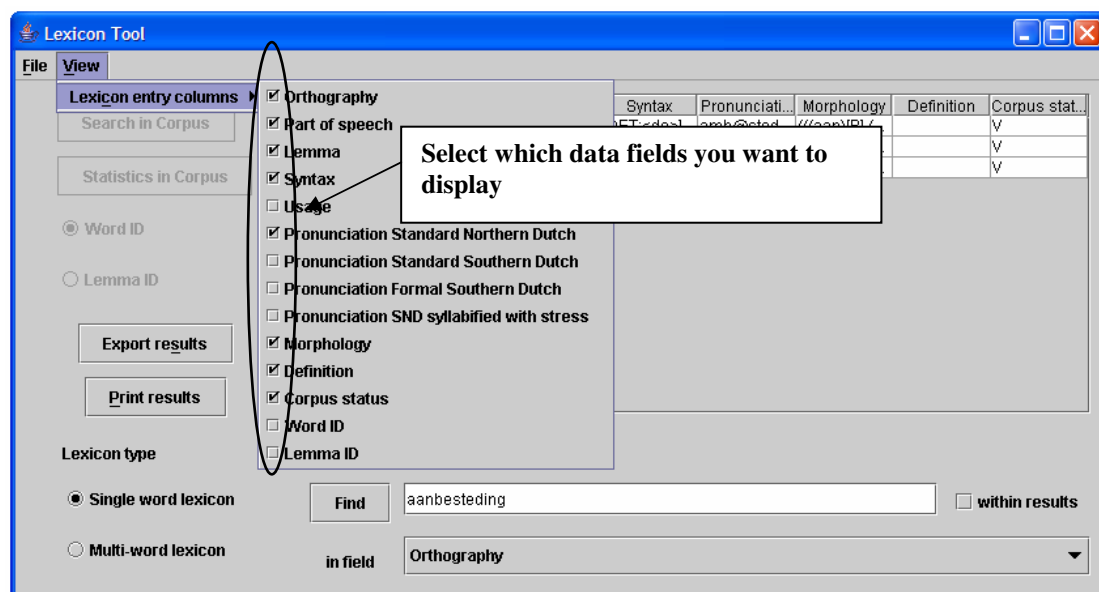




Then, if there are any matches, the search results appear.

### 6.1.1 Column visibility settings

Because the number of columns with data fields is large, not all information is visible. You can enlarge each field by positioning your mouse arrow over the division between the columns and sliding the column to the desired size. Also, you can omit some data fields in the display. Select the option **View** → **Lexicon entry columns** from the menu. In the below example three data fields concerning pronunciation are omitted as well as the fields **Usage** and **Word ID** and **Lemma ID** (the last two fields are omitted by default).

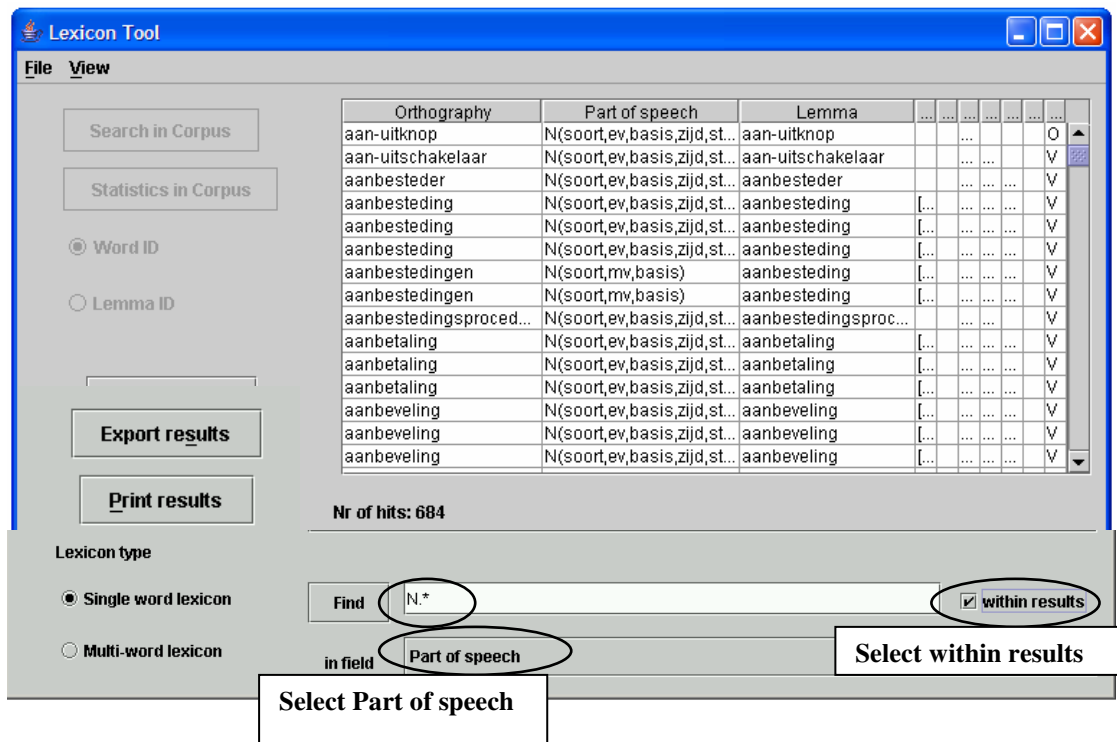


## 6.2 Searching Within Results

The lexicon tool does not allow adding new constraints to a query. However, adding constraints can be achieved in an indirect manner by performing a second (and third, etc.) query **within results**.

- Click in the box to the left of **within results**
- Select a new search criterion, using regular expression search
- (Optional) Repeat the procedure as many times as wished

The lexicon fields do not have predefined pull-down menus as do the pull-down menus in the content search panel. You have to type in the search criterion yourself. This gives considerable explorative power, however, once you master the Perl syntax. Remember the most important variables are mentioned in section [4.2](#).



In the above example a query **within results** is done. The previous selection obtained all words that begin with 'aan'. In the present query we want to extract the nouns from this selection. The way to put this in Perl syntax is `N.*`, meaning a letter string beginning with N and a subsequent row of arbitrary characters.

### 6.3 Saving and printing the search results

The selection can be saved by clicking on **Export results**. The selection is saved as a **.lex** file, i.e. the same file format as the lexicon itself (**.xml** format). It is not meant to be re-opened using Corex, but it can be opened using other applications.

The results can also be printed by clicking on **Print results**.

**! Note:** The results will be printed as they are displayed on the screen. It may be useful to limit the number of visible columns and/or adjust some column widths before printing (see 6.1.1).

### 6.4 Searching the CGN with word/lemma ID

Based on a search result in the lexicon, you can use a word or lemma as a starting point for a content search in the CGN. It is **NOT** possible to select multiple words or lemmas at once.

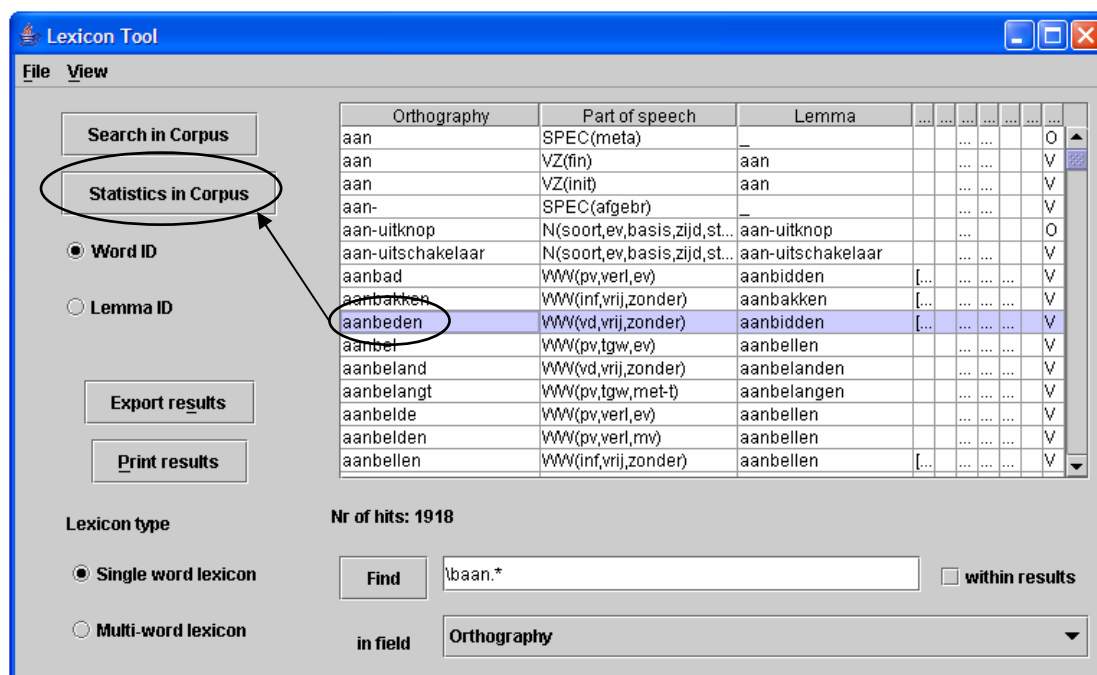
Select a word (or lemma, which amounts to the same) with the left mouse button.

- Click on the button **Search in Corpus**

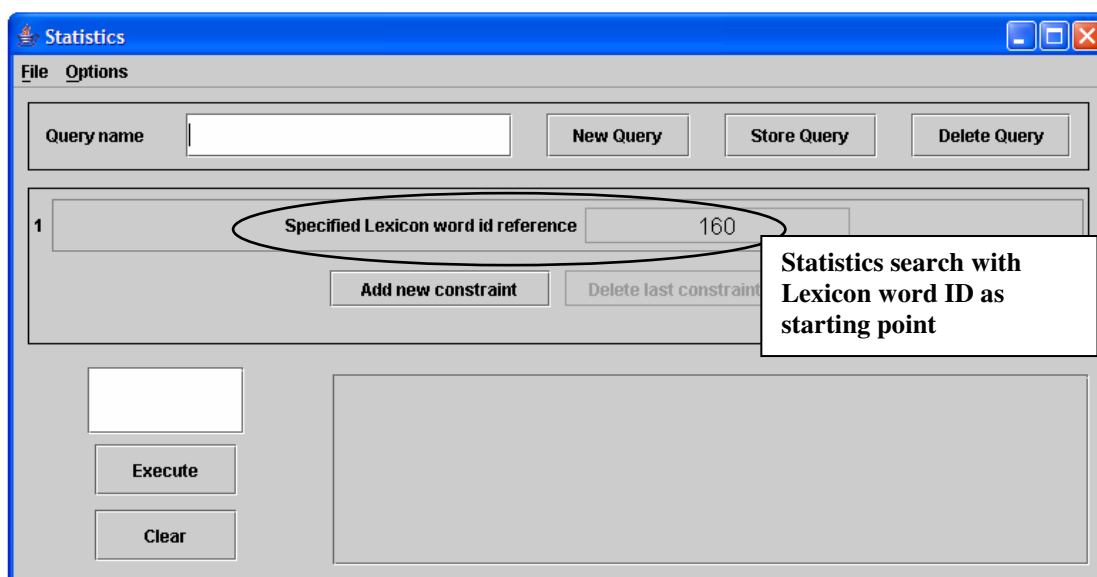


## 6.5 Statistics based on word/lemma ID

Just as in the above section, a word or lemma can be used as starting point for a Statistics Search.

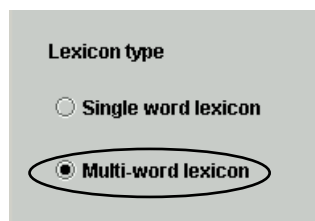


The below Statistics panel then appears. The specified lexicon ID reference is already filled in. To conduct a Statistics search, take the steps that are described in section 5. The only difference is that the **text field box** does not apply: this place is taken by the word ID or lemma ID. You can add further search conditions in the usual way, by using Add new constraint.



## 6.6 The multi-word lexicon

The multi-word lexicon is accessed by changing the default selection of the Single word lexicon to Multi-word lexicon. This option is in the bottom left area of the **lexicon panel** (this can also be done by choosing File → Lexicon Type → Multi-word lexicon.)



The use for the multi-word lexicon is finding lemmas or words that are orthographically separated but that belong to one lemma. For example: the lemma *opbellen*. The *bel op* variant of this lemma cannot be extracted from the CGN by orthographic content search without avoiding false hits (for example: “hij belde aan op een vroeg tijdstip” would be a typical false hit produced by orthographic search).

Lexicon Tool

File View

Search in Corpus

Statistics in Corpus

Multi-word lemma ID

Export results

Print results

Orthograp...	Lemma	Morphology	Definition	Continuity	Part-Ortho...	Part of sp...	Part-Optio...
bel op	opbellen	((op)[P],(bel		N	bel op	VWV(pv,tgv VZ(fin)	J J
belde op	opbellen	((op)[P],(bel		N	belde op	VWV(pv,ve VZ(fin)	J J
belden op	opbellen	((op)[P],(bel		N	belden op	VWV(pv,ve VZ(fin)	J J
bellen op	opbellen	((op)[P],(bel		N	bellen op	VWV(pv,tgv VZ(fin)	J J
belt op	opbellen	((op)[P],(bel		N	belt op	VWV(pv,tgv VZ(fin)	J J

Nr of hits: 6

Find   within results

in field

Subsequent search in the corpus for one of the hits can be done as shown in 6.3.

! Note: If performing a statistics search based on a multi-word lexicon entry, pay attention to the following. The different parts of a multi-word entry cause one hit each. For example, ‘op’ and ‘lopen’ will be counted as two occurrences of the word ‘oplopen’. So the number of hits must be divided by two. Why doesn’t Corex make this simple division? This is because if other conditions are added to the query, this could lead to incorrect results.

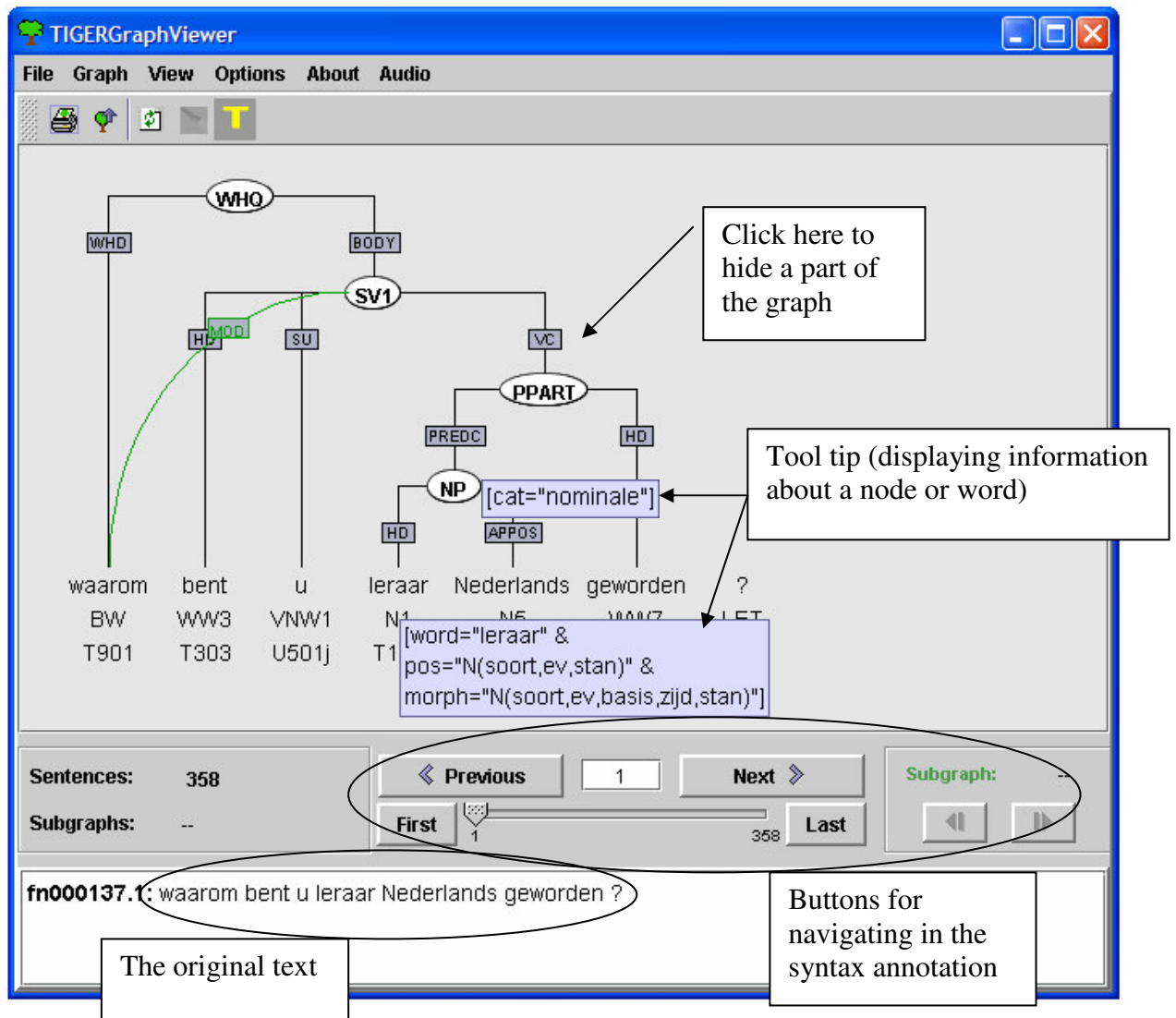
## 7 Syntax Search Panel

- About five percent of the CGN is annotated syntactically. This part of the CGN can be accessed by choosing the **Annotation types** → **syntactic annotations** option in the **Bookmarks** panel.
- The syntactic annotations can be viewed and searched by using the Tiger Syntax viewer and Tiger Search. These programs are integrated in COREX.
- A query with **Tiger** can be followed by Corex queries such as POS, but not the other way around. If you conduct a multiple-condition-query including a syntax criterion, always start with the syntax query in **Tiger**.
- The first time (and only the very first time) that you use Syntax Search, you have to load the CGN corpus. All subsequent times that you use Syntax Search, Tiger remembers the corpus that was loaded the previous time.
- A Syntax query can also be performed in a part of the CGN. This depends on the part that is selected (in the basket) prior to choosing Syntax Search.

### 7.1 The TIGERGraph Viewer

The **TIGERGraph Viewer** is accessed from the **Metadata Descriptions Tree** panel of the **Corpus Browser** window: double-click on a session node to open it, click with the left mouse button on the open node to highlight it, click with the right mouse button on the highlighted node to open the drop-down menu, then select **Tiger** from the drop-down menu (see section [1.1.2](#)). Note that this option is only available if a **.tig** (syntax) file is available.

The **TIGERGraph Viewer** is a view panel that displays syntactical information, e.g.:



The **TIGERGraph Viewer** supports the following options:

(1) Displaying the syntactical structure of each sentence in the form of a tree structure.

To view information about a node or a word, point to it with the mouse. A blue box (the so-called tool tip) will display the corresponding information.

To hide any part of the tree structure, click with the middle mouse button on a syntactic node. The corresponding structure will be hidden.

(2) Navigating in the session file by means of the following options:

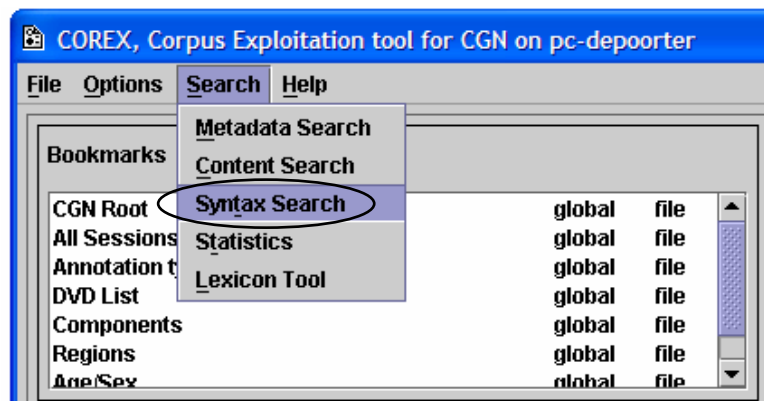
- Press **Previous** / **Next** to move to the previous / next sentence.
- Press **First** / **Last** to move to the first / last sentence.

- Enter the sentence number into the box between **Previous** and **Last** and press ENTER to jump to the corresponding sentence.
  - Click on the scroll bar to jump to a sentence.
- (3) Printing the tree structure of the displayed sentence. Make use of the menu item **Graph / Print**.
- (4) Changing the colour and display options of the **TIGERGraph Viewer**. Make use of the menu items **Options / Colour Options** and **Options / Display Options**.

For details, please see the separate manual for the “TIGERSearch” tool (a copy of the manual is on your distribution CD-ROM; you can also download the manual from <http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/manual.shtml>).

## 7.2 Syntax Search

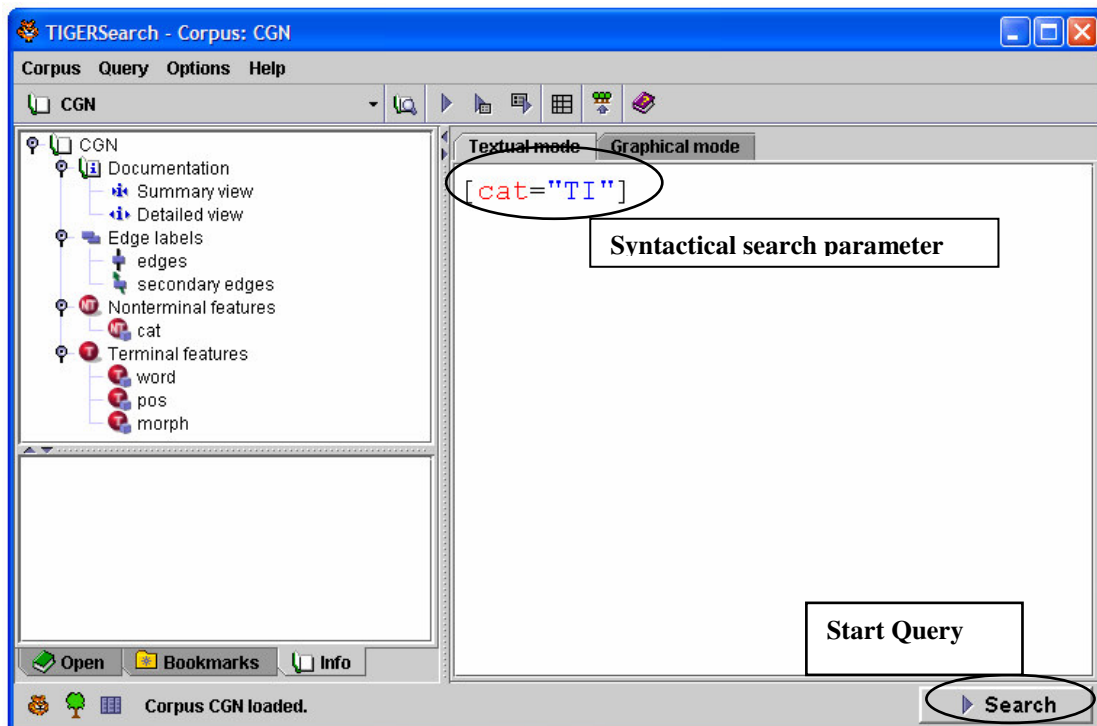
- 1) A syntax search is started by choosing the option **Syntax** from the **Search** menu. The TIGERSearch program is then opened.
- 2) TIGERSearch starts loading the dataset, which is the subset of the CGN that is annotated in such a way that syntactical searches can be performed in it.
- 3) Only the entire syntactically annotated subset of the CGN can be input for a syntax search: no subset of this subset.
- 4)



### 7.2.1 Specify the search options

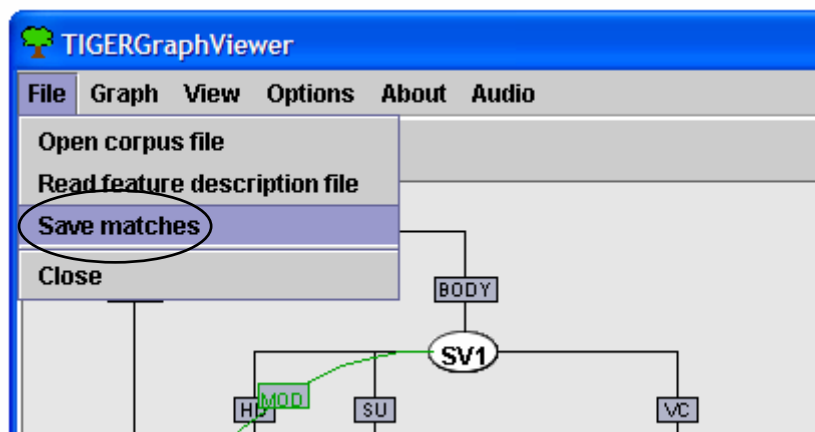
The TIGERSearch program allows syntactical search criteria to be entered, either textually or graphically. In the below example this is a query for sentences with a “te-infinitief”, codes as TI. As the query proceeds, a panel is displayed, showing the progress. When the process is finished, the TIGERGraph Viewer appears, displaying the first of the hits.





## 7.2.2 Saving the search results

The search results can be saved by choosing **Save matches** from the **File** menu of the TIGERGraphViewer.



The result is saved as a **.res** file, i.e. in the same format as is used by Content Search. This offers the possibility to reopen it in the Content Search panel of Corex and execute additional queries within this saved selection (see section [4.7](#)).

## Appendix A: example of .res file

An example of the format of the search result file (see section [4.6](#)) is as follows:

hitNr: 0

fileName: /home/broeder/COREX4/Corpora/cgn/annot//r5v1\_01/sea/fv601274.sea

tierName: V60196

trackType: ORT

utteranceNr: 2

ortContent: aan de ene kant mensen die handjes schudden groepsfoto's laten maken met fiets zonder fiets en voor de rest vergaderen.

hitWordNr: 0

hitBeginPos: 73

hitEndPos: 78

hitNr: 1

fileName: /home/broeder/COREX4/Corpora/cgn/annot//r5v1\_01/sea/fv601274.sea

tierName: V60196

trackType: ORT

utteranceNr: 2

ortContent: aan de ene kant mensen die handjes schudden groepsfoto's laten maken met fiets zonder fiets en voor de rest vergaderen.

hitWordNr: 0

hitBeginPos: 86

hitEndPos: 91

## Appendix B: CGN Metadata

N.B. Not all keys are specified for each session

### Session Keys (via Session)

CGN.wordCount	number of words in the sample: number
CGN.recCount	duration of the sample expressed in the total number of seconds: number
CGN.byteCount	indication of the size of the .wav file (expressed in terms of a number of units): number
CGN.tempoAV	average number of words per hour: number
CGN.recDate	recording date: (date or) year
CGN.locName	place where the recording was made represented in terms of the (reduced) postal code or description of place in which the recording was made; possibly unknown or unspecified.
CGN.locale	description of the type of space in which the recording was made: loc1 = room of average size; loc2 = open air; loc3 = public place; loc4 = large room; unspecified
CGN.segmentation:	the way the speech signal is synchronised with the annotation; manual or automatic
CGN.availability	label of the DVD on which the sound file can be found; e.g. CGN_WAV_01
CGN.fon.available	{true, false} whether there is a manually verified phonetic annotation available or not
CGN.syn.available	{true, false} whether there is a syntactic annotation available or not
CGN.pro.available	{true, false} whether there is a prosodic annotation available or not

### Participant Keys (via Session/Participant)

CGN.age	age class to which speaker belonged at the time the sample was recorded; age0 = under 18 years of age; age1 = 18-24 years of age; age2 = 25-34 years of age; age3 = 35-44 years of age; age4 = 45-55 years of age; age5 = over 55 years of age; ageX = age unknown
CGN.birth.year	speaker's year of birth; in case the information is not available 19nn is given as year of birth
CGN.birth.place	speaker's place of birth, represented in terms of the first three digits of the postal code preceded by the country code (B for Belgium, NL for The Netherlands; e.g. B-994, NL-832). For larger cities in the Netherlands several postal codes may apply. When this is the case, the postal code has been represented in terms of the first two digits that these codes have in common, while the variable third digit has been replaced by a hyphen (e.g. NL-25-). Where information concerning the place of birth is not available, xxx has been used. For speakers not born in

	Belgium or The Netherlands the place of birth is represented by the country code only.
CGN.birth.region	(geographical) region where the speaker was born. For a list of regions distinguished, see below.
CGN.firstLang	language (variety) speaker was raised in: SD = Standard Dutch; regiolect (e.g. regiolect: Antwerpen); dialect (e.g. dialect: Bree); unknown
CGN.homeLang	language (variety) speaker uses at home: SD = Standard Dutch; regiolect (e.g. regiolect: Antwerpen); dialect (e.g. dialect: Bree); unknown
CGN.workLang	language (variety) speaker uses at work: SD = Standard Dutch; regiolect (e.g. regiolect: Antwerpen); dialect (e.g. dialect: Bree); unknown
CGN.residence.place	speaker's place of residence, represented in terms of the first three digits of the postal code preceded by the country code (B for Belgium, NL for The Netherlands; e.g. B-994, NL-832). For larger cities in the Netherlands several postal codes may apply. When this is the case, the postal code has been represented in terms of the first two digits that these codes have in common, while the variable third digit has been replaced by a hyphen (e.g. NL-25-).
CGN.residence.reg	(geographical) region where the speaker resides. For a list of regions distinguished, see below.
CGN.residence.size	indication of the (present) size of the place where the speaker resided while (s)he was between 4 and 16 years of age; size1 = over 100,000 inhabitants; size2 = between 50,000 and 100,000 inhabitants; size3 = between 25,000 and 50,000 inhabitants; size4 = between 10,000 and 25,000 inhabitants; size5: between 5,000 and 10,000 inhabitants; size6 = fewer than 5,000 inhabitants; sizeX = size unknown
CGN.education.place	place where speaker attended secondary education place where speaker lived for the most part between ages 4 and 16 represented in terms of the first three digits of the postal code preceded by the country code (B for Belgium, NL for The Netherlands; e.g. B-994, NL-832). For larger cities in the Netherlands several postal codes may apply. When this is the case, the postal code has been represented in terms of the first two digits that these codes have in common, while the variable third digit has been replaced by a hyphen (e.g. NL-25-).
CGN.education.opleiding	type of education; e.g. lager onderwijs, mbo, universiteit
CGN.education.reg	(geographical) region where speaker lived while (s)he attended secondary education (region where speaker lived for the most part between ages 4 and 16). For a list of regions distinguished, see below.
CGN.education.level	level of education: edu1 = high, edu2 = middle, edu3 = low, eduX = unknown
CGN.occupation.level	occupational level. For the Netherlands occupational levels "occ1" up to and including "occ9" are distinguished. For Flanders, the levels "occa" up to and including "occj" are distinguished. In case someone has been trained for one

occupation but presently holds some different job, this has been indicated by combining two or more occLevel descriptions, as for example in “occC+G” where a professor is also a politician. “occX” is used whenever the occupational level is unknown.

CGN.occupation speaker’s occupation

## Content Keys (via Session/Content/Keys)

CGN.textclass.target gives information about four aspects: text type, degree of preparedness, mode and domain;  
text type specifies the component to which a sample belongs; 15 text types are distinguished; tta-tto (see list below);  
degree of preparedness: prep1 = scripted, prep2 = unscripted, prep3 = more-or-less scripted;  
mode: mod1 = broadcast, radio; mod2 = broadcast, tv; mod3 = non-broadcast;  
domain: dom1 = private; dom2= public.

CGN.textclass.keywords one or more keywords that characterize the subject matter in the sample

CGN.activity short description of activity speaker(s) was (were) involved in at the time of recording

### Text types:

tta spontaneous conversations (face-to-face)  
 ttb interviews with teachers of Dutch  
 ttc spontaneous telephone dialogues (recorded via a switchboard)  
 ttd spontaneous telephone dialogues (recorded on MD with local interface)  
 tte simulated business negotiations  
 ttf interviews/discussions/debates (broadcast)  
 ttg (political) discussions/debates/meetings (non-broadcast)  
 tth lessons recorded in a classroom  
 tti live (e.g. sport) commentaries (broadcast)  
 ttj news reports/reportages (broadcast)  
 ttk news (broadcast)  
 ttl commentaries/columns/reviews (broadcast)  
 ttm ceremonious speeches/sermons  
 ttn lectures/seminars  
 tto read speech

### Geographical regions:

regN1a The Netherlands, central region, Zuid-Holland, excl. Goeree Overflakkee  
 regN1b The Netherlands, central region, Noord-Holland, excl. West-Friesland  
 regN1c The Netherlands, central region, West-Utrecht, incl. the city of Utrecht

regN2a The Netherlands, transitional region, Zeeland, incl. Goeree Overflakke and Zeeuws-Vlaanderen  
 regN2b The Netherlands, transitional region, Oost-Utrecht, excl. stad Utrecht  
 regN2c The Netherlands, transitional region, Gelders rivierengebied, incl. Arnhem and Nijmegen  
 regN2d The Netherlands, transitional region, Veluwe up to the river IJssel  
 regN2e The Netherlands, transitional region, West-Friesland  
 regN2f The Netherlands, transitional region, Polders  
 regN3a The Netherlands, peripheral region 1 (north east), “Achterhoek”  
 regN3b The Netherlands, peripheral region 1 (north east), Overijssel  
 regN3c The Netherlands, peripheral region 1 (north east), Drenthe  
 regN3d The Netherlands, peripheral region 1 (north east), Groningen  
 regN3e The Netherlands, peripheral region 1 (north east), Friesland  
 regN4a The Netherlands, peripheral region 2 (south), Noord-Brabant  
 regN4b The Netherlands, peripheral region 2 (south), Limburg  
 regNx The Netherlands, unknown  
 regV1 Flanders, central region (Antwerpen and Vlaams-Brabant)  
 regV2 Flanders, transitional region (Oost-Vlaanderen)  
 regV3 Flanders, peripheral region 1 (West-Vlaanderen)  
 regV4 Flanders, peripheral region 2 (Limburg)  
 regVx Flanders, unknown  
 regW Wallonia  
 regZ region known to be outside of The Netherlands and Flanders  
 regX region unknown

### **Occupational level:**

#### Dutch codes

occ1 occupation requiring higher level of education (doctor, lawyer, etc.)  
 occ2 occupation requiring middle level of education (teacher, journalist, etc.)  
 occ3 occupation requiring lower level of education (mechanic, teacher nursery school, bank employee, etc.)  
 occ4 occupation not requiring any level of education (garbage collector, cleaning lady, taxi driver, etc.)  
 occ5 holding no job, unemployed  
 occ6 holding no job, attending school  
 occ7 holding no job, housewife  
 occ8 holding no job, declared unfit  
 occ9 holding no job, other

#### Flemish codes

occA occupation in higher management or government  
 occB occupation requiring higher education  
 occC employed on the teaching or research staff in a university or a college  
 occD employed in an administrative office or a service organisation  
 occE occupation not requiring any level of specification  
 occF self-employed  
 occG politicians  
 occH employed with the media (journalist, reporter) or artist  
 occI student, trainee

occJ holding no job  
occX unknown