Lexical Annotation Workbench 0.7

A Brief User Guide

1 Overview

Lexical Annotation Workbench is an integrated environment for morphological annotation. It supports simple morphological annotation (assigning a lemma and tag to a word), integration and comparison of different annotations of the same text, searching for particular word, tag etc. The workbench runs on all operating systems supporting Java, including Windows and Linux. It is an open system extensible via plugins – e.g., views, import/export filters, helps. The primary file format is the PDT 2.0 format[1] (a PML[2] instantiation); it is possible to import and export data in the PDT 1.0 format[3] (CSTS) and the TNT[4] format. In the near future, the workbench will be extended with various statistical analysis functions and assistance from external tools like taggers and parsers.

The application is organized around layers of annotation.[5] Typically, each morphological layer (an m-layer) contains information provided by a specific tool or process, e.g., a morphological analyzer, a tagger, an annotator. In the PDT 2.0 format, each layer corresponds to a single file.

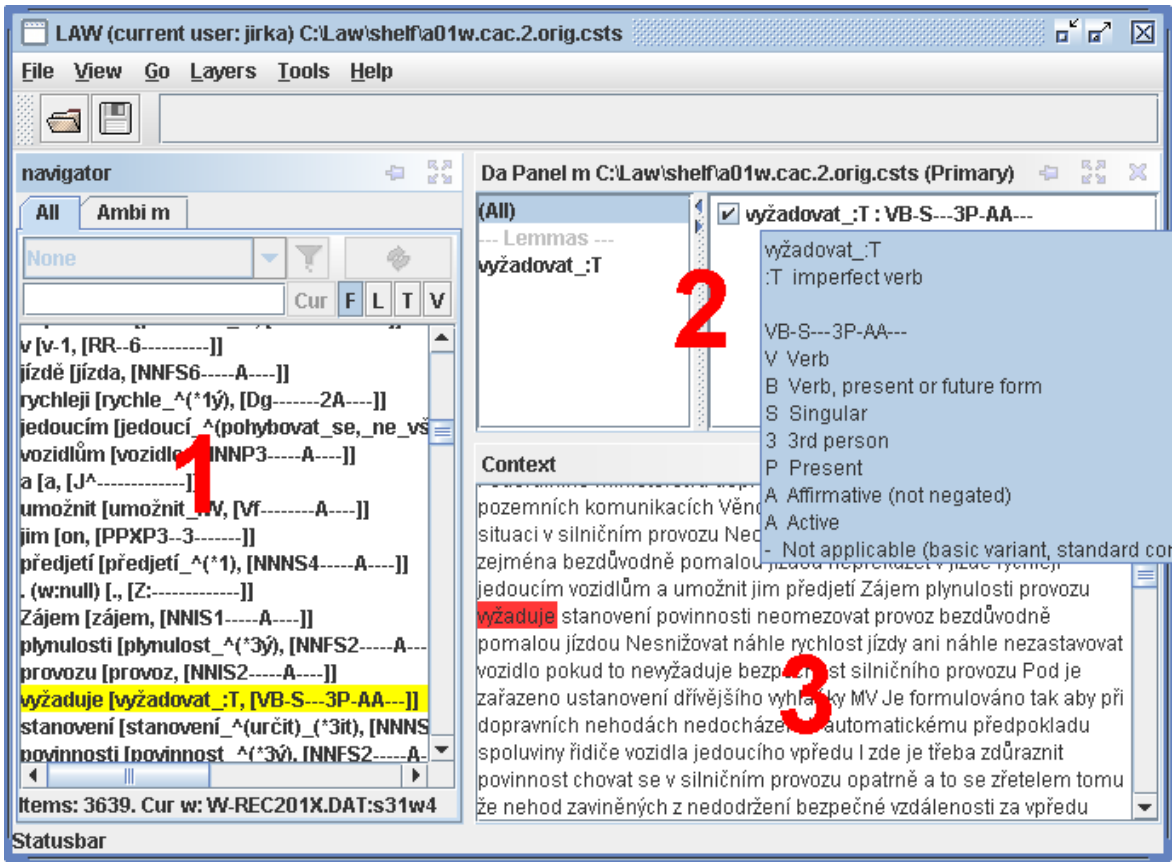
[1] See http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch03.html#a-data-formats-pml
[2] http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/pml/index.html
[3] http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch03.html#a-data-formats-csts
[4] http://www.coli.uni-saarland.de/~thorsten/tnt/
[5] See http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch02.html.

1.1 Major components

As can be seen on the screenshot below, the application consists of three major components:

1. **Navigator** (see §2below) – for navigating through forms of the document (browsing, filtering, searching, etc.)

2. **DaPanels** (see §3below) – for displaying and disambiguating morphological information (lemmas, tags) of a form. One of the panels is primary
3. **Context Views** – various contexts, e.g. plain text of the document, syntactic structures, etc. The default and most often used context view is DocView which simply displays the processed text.



1.2 The Usual Process of Disambiguation

The usual disambiguation proceeds as follows.

1. Open a file with the m-layer you want to annotate (`File > Open`). The corresponding w-layer will be opened automatically. You may open multiple m-layers referring to a single w-layer. For example, you may want to open an m-layer produced by a morphological analyzer and two m-layers, each produced by a different tagger.
You may also import layers from the csts or tnt format. In that case, do not forget to specify the correct character encoding in the `Open` dialog (e.g. ISO Latin 2 for PDT 1.0).

2. If you opened or imported multiple m-layers most likely you will annotate just one of them. Select the DaPanel of that layer as primary (`View > Choose Primary DaPanel`).

3. In the Navigator, switch to the `Ambi` word-list displaying the ambiguous words of the layer you want to annotate and select the first word.

4. Press `Enter`. The cursor moves to the primary DaPanel. Select the correct lemma and tag and press `Enter` again. The cursor will move to the next ambiguous word.

If the correct lemma and/or tag are not offered you can insert them by pressing `Ins` or selecting `Insert` from the context menu (right-click within the DaPanel to display the context menu).

If you made a mistake, switch to the `All` word-list, find the word where you made the mistake, select it, the associated annotation appears in the DaPanel. You can now select the correct item and then switch back to the `Ambi` word-list.

5. Save the result of your work by `File > Save` command. If you want to save only the selected lemmas and tags delete all unselected items by `Layers > Remove Unselected Items` command before saving the layer.

1.3 Commands

File

- `Open ...` – Opens a layer. The layer must be compatible with the currently loaded w-layer (if there is one). If you want to open m-layers referring to a different w-layer, close the currently opened layers first.

- `Save` – Saves the currently opened layers

- `Export` – Export currently opened layers to a non-PDT 2.0 format (currently CSTS or TNT).

- `Close` – Closes the currently opened layers (suggesting to save possible changes)

- `Exit` – Exits the application

View

- `Choose Primary DaPanel ...` – Enables you to change the primary panel
- `New DaPanel ...` – Opens a new DaPanel.
- `Views ...` – Displays a list of all open views

Go

- `Go To Navigator` – Moves focus to the list of words of the Navigator.
- `Go To Groups` – Moves focus to the list of groups (tags grouped by lemmas, POS, etc) of the primary DaPanel.
- `Go To Items` – Moves focus to the list of items of the primary DaPanel.
- `Go There` – Moves focus from navigator to groups, from groups to items.
- `Go Back` – Moves focus from items to groups, from groups to the navigator.
- `Go Auto` – Selects the current item and moves focus to the next choice point.

Layers

- `Duplicate a Layer ...` – Creates a copy of an existing layer.
- `Copy Selection ...` – Copies the selection from one layer to another.
- `Remove Unselected Items ...` – Removes all unselected items from a layer
- `Compare Layers ...` – Creates a navigator list containing items where two layers differ.
- `Info ...` Displays basic information about a layer

Tools

- `Save Settings` – Saves current settings (e.g. saved filters, Open file directory, etc.). The settings are also saved when the application exists.

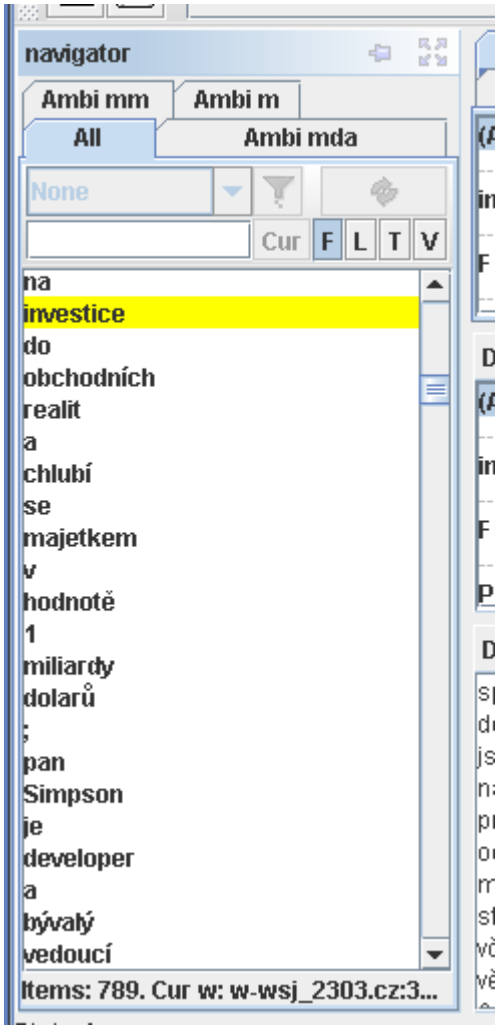
Help

- `Help ...` – Currently not implemented.
- `About ...` – Displays info about the application and its version.

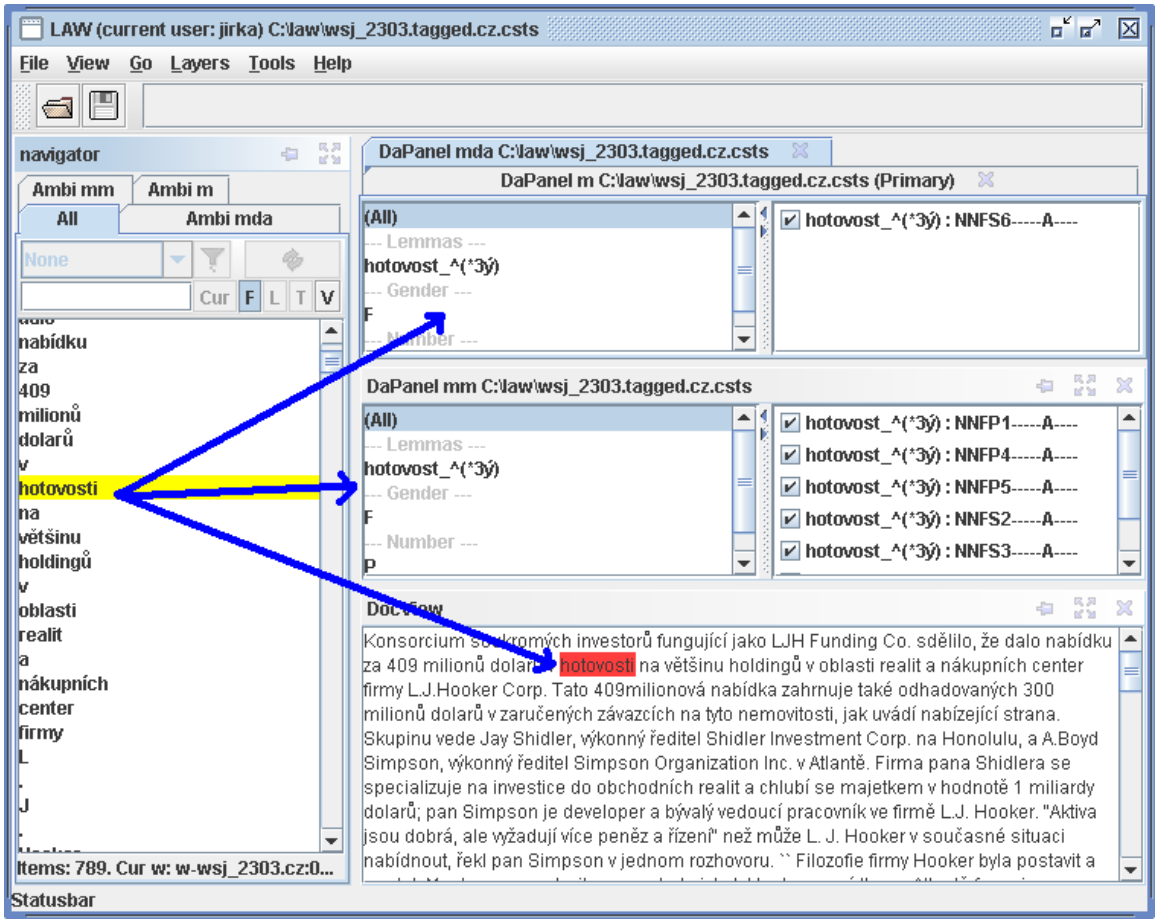
2 Navigator

2.1 Overview

Navigator is the main component for accessing words of the document and the morphological information associated with them. It consists of several lists of words (word-lists).



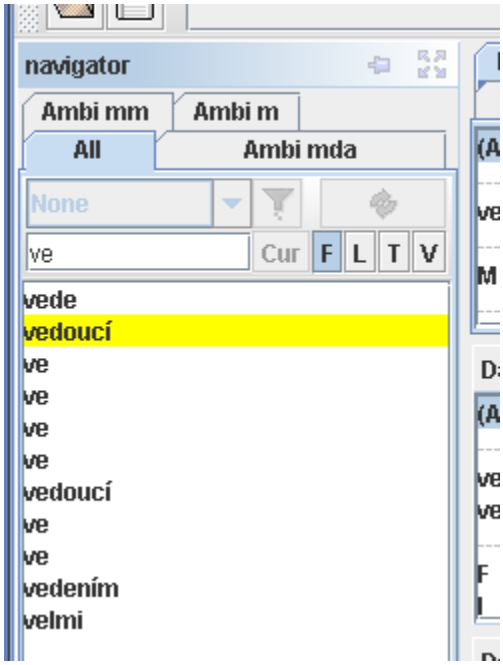
The `All` word-list contains all the words of the document (if not filtered, see below). An `Ambi` word-list contains all the words ambiguous in a particular layer. If you select a word in the navigator, the DaPanels will display the word's lemmas and tags and the context views will display the word's context.



You can filter the words by their form, lemmas or tags – see 2.2 Filtering below.

The words are always displayed. You can also display associated lemmas and/or tags of a particular layer, either all, or only the selected ones. See 2.3 Configuration below.

2.2 Filtering



- You can filter by specifying the desired forms, lemmas or tags. You can consider only selected or all lemmas/tags. If you opened more than one m-layr and want to configure which layer is considered in filtering, use the `Change Current Layer` command, see 2.3 Configuration below.

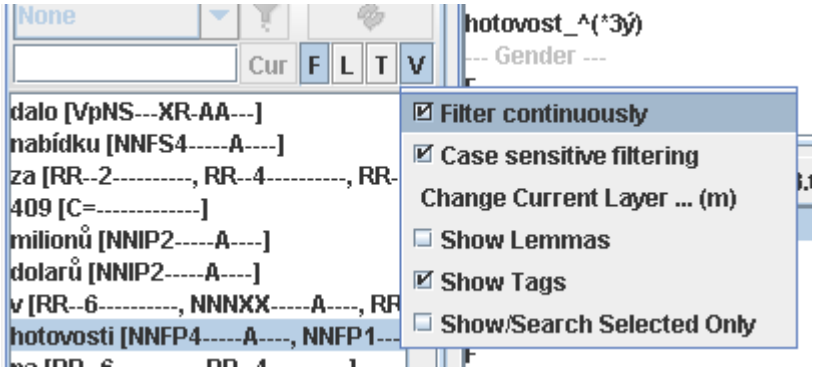
- You can use regular expressions; see 2.4 Filtering with regular expressions below.

- All filters are considered to be prefixes. If you want to match whole words add `$` at the end. For example, `v` matches *v, ve, venku* etc., while `v$` matches only *v*.

- You can save your filter for future use by pressing `Ctrl + Enter`.

- You do not need to display lemmas and/or tags in the Navigator to filter by them.

2.3 Configuration



- `Filter continuously` – the filtering is applied immediately as you type the filter. Otherwise you have to press Ctrl + Enter to apply the filter. The continuous filtering can be slow for large files.

- `Case sensitive filtering` – should the filter consider case of letters or not.

- `Change Current Layer ...` – Change the filtering layer of this list. The current layer is considered when you filter by lemmas or tag, or when navigator shows lemmas and/or tags.
  By default, the first opened m-layer is the default current layer in the `All` list. In an `Ambi` list, the default current layer is the same as the m-layer that determines whether a word is ambiguous. You can change these default settings using the command `Change Current Layer`

- `Show Lemmas` – For each word, its lemmas in the current layer are displayed.

- `Show Tags` – For each word, its tags in the current layer are displayed.

- `Show/Search Selected Only` – Only selected items (lemmas and/or tags) are displayed. Also, filtering considers only selected items.

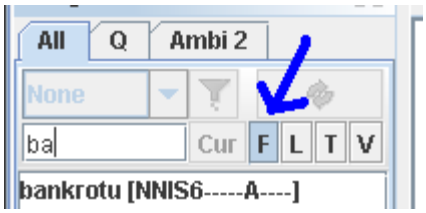
2.4 Filtering with regular expressions

Note: Every expression is considered to be a prefix, i.e. `.*` is automatically appended. Add `$` at the end of the item if you want to avoid this (`x$` is the same thing as `x$.*`)

2.4.1 Examples

Filtering forms

The button `F` next to the filter field must be depressed:



ba	words beginning with *ba*
.*ba$	words ending in *ba*
ba$	words *ba*
\d+$	all numbers

Filtering lemmas

The button `L` next to the filter field must be depressed.

.*_;[RK]	product or company names (assuming the PDT term classification[6])
.*_,t	foreign words (assuming the PDT style classification[7])
.*_;[RK]_,t	foreign product or company names

Filtering tags

The button `T` next to the filter field must be depressed.

These examples assume the Czech Positional Tagset used in PDT.

N	nouns
[^N]	not nouns
NN[MF]	masculine and feminine nouns
NN.[1-4]	nouns in nominative, genitive, dative and accusative
NN.[^1]	nouns not in nominative
NN(FS[46]|MP[^1])	nouns that are either feminine singular accusative/locative or masculine plural not in nominative
.{14}[^-]	words not in their basic variant
.*[^-]$	words not in their basic variant
.{14}6	words not in the variant `6`

2.4.2 Syntax

This is only a simplified overview of the most common constructs. For a more detailed description of regular expression constructs, see Jeffrey E. F. Friedl: Mastering Regular Expressions, 2nd Edition, O'Reilly and Associates, 2002.

[6] http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/ch02s01s04.html
[7] http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/ch02s01s05.html

Character classes

.	Any character
[abc]	a, b, or c
[^abc]	Any character except a, b, or c (negation)
[a-zA-Z]	a through z or A through Z, inclusive (range)
\d	A digit: [0-9]
\D	A non-digit: [^0-9]
\p{L}	A letter (note that [a-zA-Z] excludes letters with diacritics, etc)
\p{Lu}	An uppercase letter
\p{Punct}	Punctuation: One of !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~
\p{InGreek}	Greek letters
\P{X}	Any character except X, e.g. \P{L} is any character that is not a letter

Boundaries

^	The beginning of an item
$	The end of an item

Quantifiers

X?	X, once or not at all
X*	X, zero or more times
X+	X, one or more times
X{n}	X, exactly n times
X{n,}	X, at least n times
X{n,m}	X, at least n but not more than m times

These quantifiers are greedy, the corresponding reluctant (e.g. X??, X*?, X+?) and possessive quantifiers (e.g. X?+, X*+, X++) are also supported.

Logical operators

XY	X followed by Y
X|Y	Either X or Y
(X)	X, as a capturing group

Back references

\n	Whatever the nth capturing group matched

Quotation

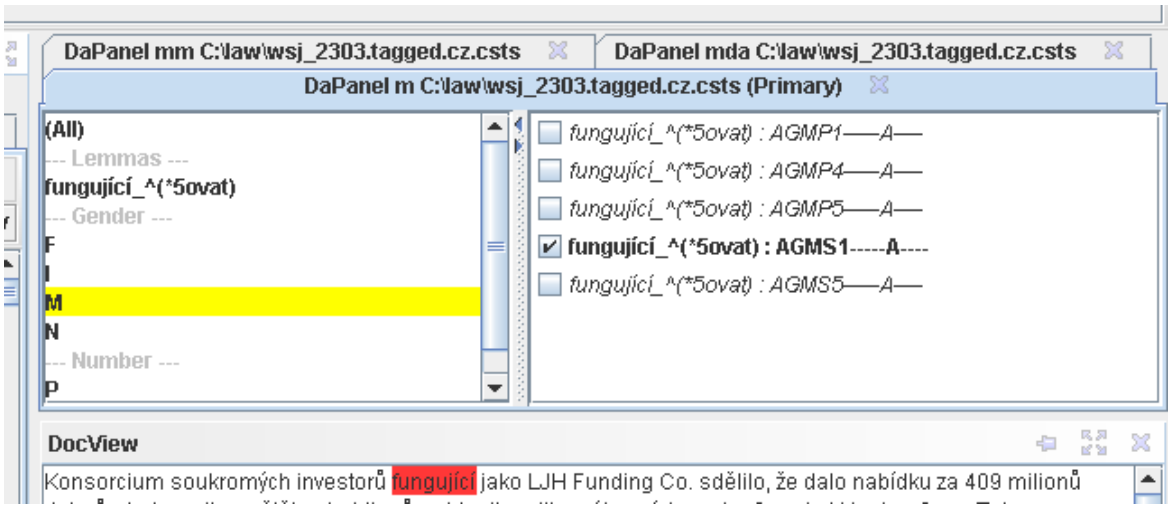
\	Nothing, but quotes the following character (e.g. \\ stands for the backslash character)
\Q	Nothing, but quotes all characters until \E (e.g. \Q\\E is equivalent to \\)
\E	Nothing, but ends quoting started by \Q

Special constructs (non-capturing)

(?:X)	X, as a non-capturing group
(?i)	Filtering is case-insensitive
(?-i)	Filtering is case-sensitive
(?i:X)	X, with case insensitive filtering (similarly -i)
(?=X)	X, via zero-width positive lookahead
(?!X)	X, via zero-width negative lookahead
(?<=X)	X, via zero-width positive lookbehind
(?<!X)	X, via zero-width negative lookbehind

3 DaPanel

DaPanel is a component used for disambiguating morphological information assigned to a word. It has two main parts – an Item List (or simply Items) on the right and a Group List (Groups) on the left. The Item List contains items (lemma and tag pairs) of the currently selected word in a particular layer. Each item can be selected or deselected. The Group List groups the items by various categories – by lemmas, POS, etc. If you select a particular group, only the items belonging to that group are displayed in the Item List. The set of group categories can be changed or the Group List can be hidden all together. The active window is marked by a yellow selection bar.



3.1 Actions

Typically, when disambiguating, one selects the appropriate item and moves to the next ambiguous word using the so-called Auto mode (pressing `Enter`). Auto mode skips words with only one item and skips the Groups if there is only one group in each category (e.g. only one lemma).

- Select/Deselect item (lemma and tag pair):
  - Mouse: click the check box next to it.

    - Keyboard: `Space`
- Go to Group List (i.e. make it active):
    - Menu: `Go > Go to Groups`
    - Mouse: click within the Group List
    - Keyboard: `Ctrl + L`
- Go to Item List (i.e. make it active):
    - Menu: `Go > Go to Items`
    - Mouse: click within the Item List (avoid the checkbox if you do not want to change selection)
    - Keyboard: `Ctrl + T`
- Move there: moves focus one level down: from the Navigator to the Groups, from the Group to the Items
    - Menu: `Go > Go There`
    - Keyboard: `Ctrl + Right`
- Move back: Moves focus one level up: from the Items to the Groups, from the Groups to the Items
    - Menu: `Go > Go Back`
    - Keyboard: `Ctrl + Left`
- Move within a list: `Up/Down`, `Page Up/Down`, `End/Home`, pressing the first letter
- Auto: combines selection/deselection actions with the most likely Go To actions. It selects the current item (deselects all others) and moves to the next choice point. Auto mode skips words with only one item and skips the Groups if there is only one group in each category (e.g. only one lemma).
    - Menu: `Go > Go Auto`
    - Keyboard: `Enter`
- Display item help – help explaining tag values and lemma terms
    - Mouse: Leave the mouse cursor over an item for a while
    - Keyboard: `F1` or `Ctrl + F1`

3.2 Multiple DaPanels, Primary DaPanel

Although usually only one layer is disambiguated, any layer can be assigned a DaPanel. In such a case, one of them is primary. This panel is marked by `(Primary)` in its title bar. The Go To commands and associated shortcuts work with this primary panel. Change the primary panel by `View > Choose Primary DaPanel`.