Faculty of Electrical Engineering and Computing
University of Zagreb

TermeX v1.0

Users manual

Autor: Davor Delač

January 21, 2009

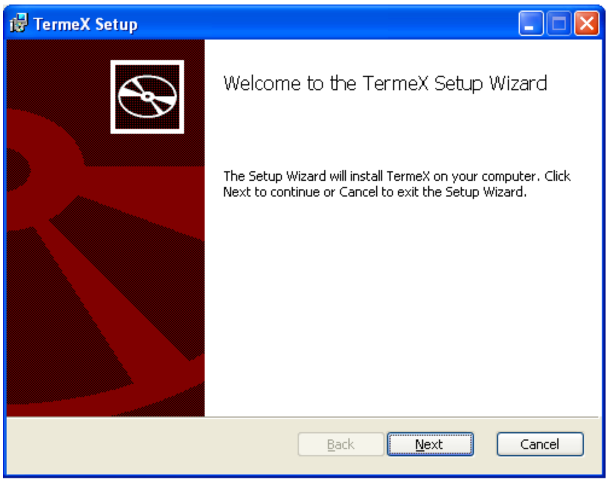
Contents

1 Introduction

TermeX is a tool for automatic collocation extraction and terminology lexica construction. It is based on statistical measures called association measures (AMs). Fourteen AMs are implemented in *TermeX*, which, combined with lemmatization, enable users faster and better construction of terminology lexica. Extraction of n-gram collocations up to length four is allowed.

2 Installation instruction for Microsoft windows

TermeX has a front end GUI designed in Java. It is necessary to acquire proper Java Runtime Environment from `http://java.sun.com/javase/downloads/index.jsp`. This application is designed using Java 1.6, therefore this version of JRE is preferred. Once the JRE is installed, installation process for *TermeX* is initiated by starting `TermeX.msi`. First window of installation process is depicted by Figure 1.



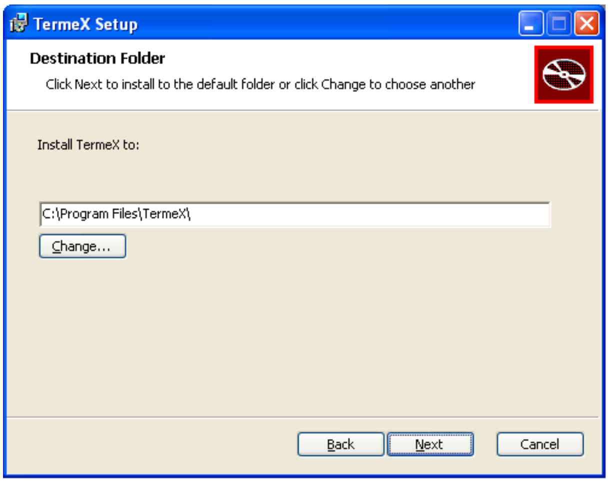
TermeX Setup

Welcome to the TermeX Setup Wizard

The Setup Wizard will install TermeX on your computer. Click Next to continue or Cancel to exit the Setup Wizard.

Figure 1: Installation start.

By selecting the option Next, the second window, given by Figure 2, is displayed. This window enables the user to select the location on disc where *TermeX* is to be installed. Further selection of the option Next finalizes the installation process.



TermeX Setup

Destination Folder

Click Next to install to the default folder or click Change to choose another

Install TermeX to:

C:\Program Files\TermeX\

Figure 2: Izbor lokacije na disku.

3 User manual

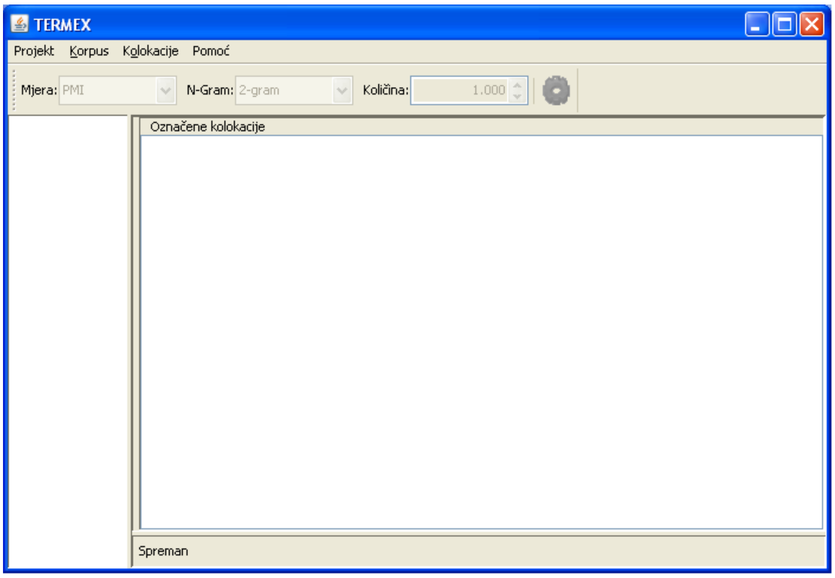
TermeX has a simple and intuitive graphical user interface. Processing of corpora is divided in project. A project represents a construction of one terminology lexicon, and can consist of several processed corpora. *TermeX* provides a wide spectrum of functionality described in following subsections.

3.1 Graphical user interface

Graphical user interface allows for fast and simple access to functionality provided by the *TermeX* tool. The interface consists of following parts:

1. *Project Explorer* is a list of currently open corpora which are being processed as a part of a project.;
2. *Working Area* is used for hand processing of corpora. It consists of tables representing query results and allowing for selection of collocations for desired lexicon.;
3. *List of selected collocations* is a list of n-grams selected by the user as a part of a future terminology lexicon.;
4. *Status Bar* displays current status of automatic corpora processing.;
5. *Toolbar* is used for fast query invocations (calculation of one AM).;
6. *Menu*.

Menu consists of the following parts:



Mjera: PMI

N-Gram: 2-gram

Količina: 1.000

Figure 3: Graphical user interface.

- Project – access to functions for manipulating with projects.
- Corpus – access to functions for adding and removing of corpora in the current project.
- Collocations – methods for faster and easier corpora processing;
- Helpć.

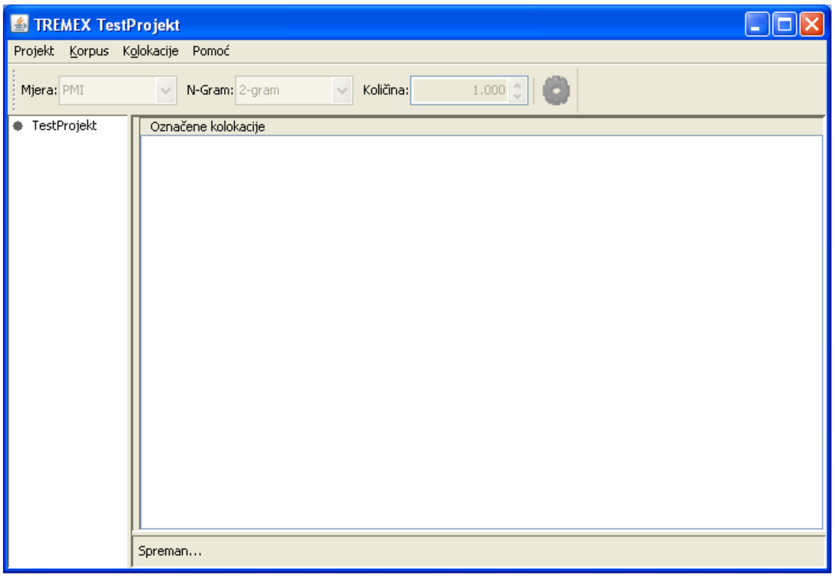
3.2 Projects

Project represents processing of multiple corpora with the purpose of creating one terminology lexicon. Menu offers access to functions for manipulating with projects:

- New Project – creates a new project. After this option is selected, a new project with the name `untitled` is created.;
- Save – saves the changes made to the project. If the project is being saved for the first time, this option resembles the “Save as...” option.;
- Save as... – used for selecting of desired location on disc where the project info is to be saved.;
- Open – opens the previously saved project;

- Close – closes the project.;

For creating of new projects, user must first select the option New Project. Name of the project is determined when the option Save As is selected for the first time. As a result, the project explorer displays the new project name.



Mjera: PMI

N-Gram: 2-gram

Količina: 1.000

Spreman...

Figure 4: New Project.

3.3 Processing corpora

Processing of a corpus begins by adding a corpus to the project. A corpus is a text file encoded in UTF-8 format. Corpora are added using Add Corous option in the Corpus menu. Upon selecting this option, a window for selecting the corpus to be added is displayed. Once the user has selected the desired corpora, it is added to the project explorer and the status of automatic processing is displayed in the status bar.

Manual processing starts by invoking a query on a corpus. A query represents calculation of AM values for n-grams of certain length and sorting of the results by Am value in decreasing order. Before a query can be invoked it is necessary to select a corpus on which the query will be invoked. This is done by double-clicking the desired corpus in the project explorer. A change of icon in project explored indicates that a corpus has been selected.

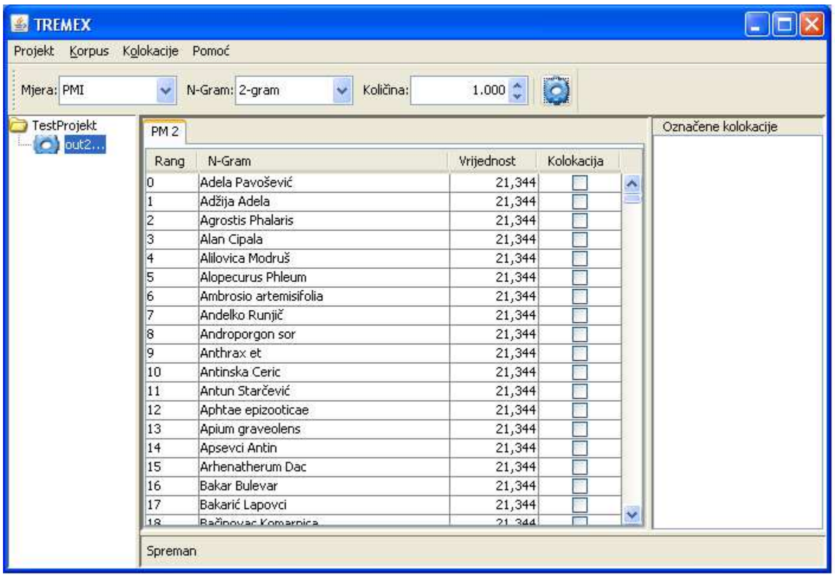
Queries are invoked using the query toolbar. This toolbar consists of four parts:

- AM – for selection of association measure (AM);

- N-gram – for selection of the desired length of n-grams;
- Amount – number of best ranked n-gram collocation candidates to be displayed in respect to selected AM;
- Execute button – invokes selected query on selected corpus.

Queries result in a table containing data on best ranked collocations in respect to selected AM. This table consists of four columns:

- Rank – rank of the collocation in a sorted list;
- N-Gram – collocation text;
- Value – value of the AM for the collocation;
- Collocation – a check box which when selected indicates that the n-gram is a collocation and a part of a future terminology lexicon.



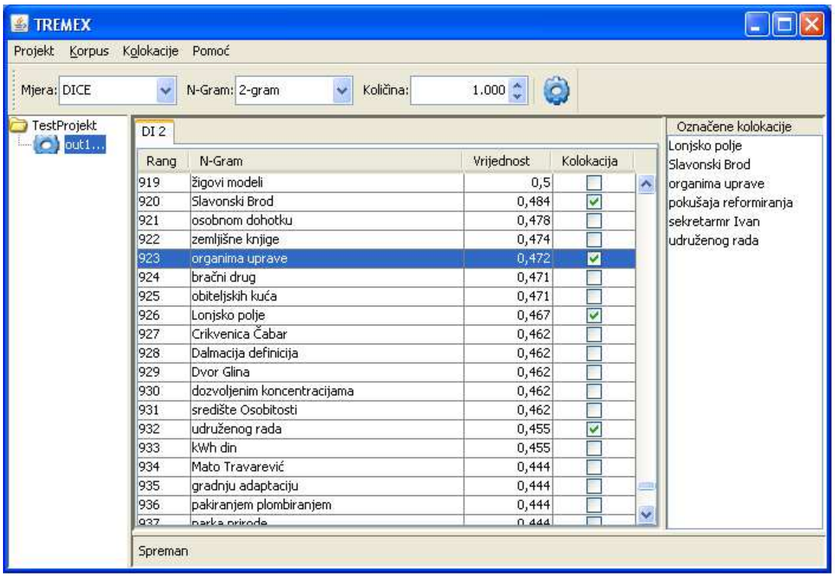
Mjera: PMI
N-Gram: 2-gram
Količina: 1.000

Rang	N-Gram	Vrijednost	Kolokacija
0	Adela Pavošević	21,344	□
1	Adžija Adela	21,344	□
2	Agrostis Phalaris	21,344	□
3	Alan Cipala	21,344	□
4	Alilovica Modruš	21,344	□
5	Alopecurus Phleum	21,344	□
6	Ambrosio artemisifolia	21,344	□
7	Andelko Runjič	21,344	□
8	Andropորgon sor	21,344	□
9	Anthrax et	21,344	□
10	Antinska Ceric	21,344	□
11	Antun Starčević	21,344	□
12	Aphtae epizooticae	21,344	□
13	Apium graveolens	21,344	□
14	Apsevci Antin	21,344	□
15	Arhenatherum Dac	21,344	□
16	Bakar Bulevar	21,344	□
17	Bakarić Lapovci	21,344	□
18	Račinovac Komarnica	21,344	□

Figure 5: Query result.

By checking the "Collocation" check box, the user indicates that the selected n-gram is a collocation. *TermeX* ensures that this collocation is selected in all existing tables and that the entry is added to the List of selected n-grams. This consistency is followed when the user decides to remove a collocation from the List of selected n-rams or to uncheck the check box in one of the tables.

When manually processing a corpus, the user has the following functions at his disposal (functions of the Collocations menu);



Mjera: DICE N-Gram: 2-gram Količina: 1.000

Rang	N-Gram	Vrijednost	Kolokacija
919	žigovi modeli	0,5	□
920	Slavonski Brod	0,484	☑
921	osobnom dohotku	0,478	□
922	zemljišne knjige	0,474	□
923	organima uprave	0,472	☑
924	bračni drug	0,471	□
925	obiteljskih kuća	0,471	□
926	Lonjsko polje	0,467	☑
927	Crikvenica Čabar	0,462	□
928	Dalmacija definicija	0,462	□
929	Dvor Glina	0,462	□
930	dozvoljenim koncentracijama	0,462	□
931	središte Osobitosti	0,462	□
932	udruženog rada	0,455	☑
933	kWh din	0,455	□
934	Mato Travarević	0,444	□
935	gradnju adaptaciju	0,444	□
936	pakiranjem plombiranjem	0,444	□
937	parka prirode	0,444	□

Označene kolokacije
Lonjsko polje
Slavonski Brod
organima uprave
pokušaja reformiranja
sekretarmr Ivan
udruženog rada

Spreman

Figure 6: Corpus processing.

- Import – import the number of selected collocations from the formatted text file. Each collocation is considered to be in the new row in the text file.;

- Export – exports a desired list of n-grams formatted so that every n-gram is in a new line.;

- Select all – selects all n-grams of the current table as collocations;

- Unselect all – unselects all n-grams of the current table.;

- Concordances – opens a window with concordances for selected n-gram. This new window consists of a table containing prefix, suffix, and byte offset from the beginning of the corpus for the desired n-ram.

4 Appendix: Association measures in *TermeX*

$$G_0(I, w_1 \cdots w_n) = \log_2 \frac{P(w_1 \cdots w_n)}{\prod_{i=1}^{n} P(w_i)}, \tag{1}$$

$$G_0(DICE, w_1 \cdots w_n) = \frac{nf(w_1 \cdots w_n)}{\sum_{i=1}^{n} f(w_i)}, \tag{2}$$

$$G_1(g, w_1 \cdots w_n) = \frac{g(w_1, w_2 \cdots w_n) + g(w_1 \cdots w_{n-1}, w_n)}{2}, \tag{3}$$

$$G_2(g, w_1 \cdots w_n) = \frac{g(w_1 \cdots w_{\lfloor n/2 \rfloor}, w_{\lceil n/2 \rceil} \cdots w_n) + g(w_1 \cdots w_{\lfloor n/2+1 \rfloor}, w_{\lceil n/2+1 \rceil} \cdots w_n)}{2}, \tag{4}$$

$$G_3(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_i, w_{i+1}), \tag{5}$$

$$G_4(g, w_1 \cdots w_n) = g(w_1 \cdots w_{n-1}, w_2 \cdots w_n), \tag{6}$$

$$G_5(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_1 \cdots w_i, w_{i+1} \cdots w_n), \tag{7}$$

$$G_6(g, w_1 \cdots w_n) = G_0\Big(g, (w_1 w_2, w_2 w_3, \ldots, w_{n-1} w_n)\Big), \tag{8}$$

$$G^* : \Omega \times W^+ \times S \to \mathbb{R}, \quad S \subset W. \tag{9}$$

$$G_0^*(I, w_1 w_2 w_3, \{w_2\}) = \log_2 \frac{P(w_1 w_2 w_3)}{P(w_1) P(w_3)}, \tag{10}$$

$$H(g, w_1 w_2 w_3) = \begin{cases} \alpha_1 G_0^*(g, w_1 w_2 w_3, \{w_2\}) & \text{if } \mathit{stop}(w_2) \\ \alpha_2 G_{4,6}^*(g, w_1 w_2 w_3, \varnothing) & \text{otherwise}, \end{cases} \tag{11}$$

$$\alpha_1 = \frac{1}{\max\limits_{\{w_1 w_2 w_3 \in W^3 | \mathit{stop}(w_2)\}} G_0^*(g, w_1 w_2 w_3, \{w_2\})}, \tag{12}$$

$$\alpha_2 = \frac{1}{\max\limits_{\{w_1 w_2 w_3 \in W^3 | \mathit{stop}(w_2)\}} G_{4,6}^*(g, w_1 w_2 w_3, \varnothing)} . \tag{13}$$

$$H_1(g, w_1 w_2 w_3 w_4) = \begin{cases} \alpha_1 G_0^*(g, w_1 w_2 w_3 w_4, \{w_2, w_3\}) & \text{if } \mathit{stop}(w_2) \wedge \mathit{stop}(w_3) \\ \alpha_2 G_0^*(g, w_1 w_2 w_3 w_4, \{w_2\}) & \text{if } \mathit{stop}(w_2) \wedge \neg \mathit{stop}(w_3) \\ \alpha_3 G_0^*(g, w_1 w_2 w_3 w_4, \{w_3\}) & \text{if } \mathit{stop}(w_3) \wedge \neg \mathit{stop}(w_2) \\ \alpha_4 G_6^*(g, w_1 w_2 w_3 w_4, \varnothing) & \text{otherwise}, \end{cases} \tag{14}$$

$$H_2(g, w_1w_2w_3w_4) = \begin{cases} \alpha_1 G_0^*(g, w_1w_2w_3w_4, \{w_2, w_3\}) & \text{if } \mathit{stop}(w_2) \wedge \mathit{stop}(w_3) \\ \alpha_2 G_0^*(g, w_1w_2w_3w_4, \{w_1, w_2\}) & \text{if } \mathit{stop}(w_2) \wedge \neg \mathit{stop}(w_3) \\ \alpha_3 G_0^*(g, w_1w_2w_3w_4, \{w_3, w_4\}) & \text{if } \mathit{stop}(w_3) \wedge \neg \mathit{stop}(w_2) \\ \alpha_4 G_6^*(g, w_1w_2w_3w_4, \varnothing) & \text{otherwise,} \end{cases} \tag{15}$$

$$H_3(g, w_1w_2w_3w_4) = \begin{cases} \alpha_1 G_0^*(g, w_1w_2w_3w_4, \{w_2, w_3\}) & \text{if } \mathit{stop}(w_2) \wedge \mathit{stop}(w_3) \\ \alpha_2 G_0^*(g, w_1w_2w_3w_4, \{w_2, w_3\}) & \text{if } \mathit{stop}(w_2) \wedge \neg \mathit{stop}(w_3) \\ \alpha_3 G_0^*(g, w_1w_2w_3w_4, \{w_1, w_3\}) & \text{if } \mathit{stop}(w_3) \wedge \neg \mathit{stop}(w_2) \\ \alpha_4 G_6^*(g, w_1w_2w_3w_4, \varnothing) & \text{otherwise,} \end{cases} \tag{16}$$