



**User Manual
crossMining**

(Last update July 22, 2010)

Copyright 2010 Across Systems GmbH

The contents of this document may not be copied or made available to third parties in any other way without the written permission of Across Systems GmbH. Though utmost care has been taken to ensure the correctness of the content, neither Across Systems GmbH nor the author assume any responsibility for errors or missing content in this document or incorrect interpretation of the content. All mentioned brands are property of the respective owners.

Table of Contents

1	Introduction	3
1.1	ABOUT CROSSMINING	3
1.2	ABOUT THIS DOCUMENTATION	4
1.2.1	<i>Icons</i>	4
1.2.2	<i>Conventions</i>	4
1.2.3	<i>Cross-reference</i>	5
1.2.4	<i>Feedback</i>	5
1.2.5	<i>Document Versions</i>	5
2	Installation	6
2.1	SYSTEM REQUIREMENTS.....	6
2.2	INSTALLING CROSSAPI INTERACTIVE	6
2.2.1	<i>Creating a Generic Softkey</i>	9
	Saving the Generic Softkey to a Storage Medium	10
	Sending the Generic Softkey by E-Mail.....	11
2.3	INSTALLING CROSSMINING.....	12
3	Using crossMining	15
3.1	OVERVIEW OF CROSSMINING FUNCTIONS.....	15
3.2	STARTING CROSSMINING.....	15
3.3	WORKING WITH CROSSMINING.....	17
3.3.1	<i>crossMining Toolbar</i>	17
3.3.2	<i>Statistical Lexica</i>	17
	Creating Statistical Lexica.....	18
	Deployment of Statistical Lexica	21
3.3.3	<i>Autocompletion</i>	21
	Autocompletion Function in Across	21
	Autocompletion Test	22
3.3.4	<i>Terminology Harvesting</i>	24
	Addition of Target-Language Terms	24
	Bilingual Term Extraction	27
3.3.5	<i>Import of Moses SMT Phrase Tables</i>	30
3.4	CLOSING CROSSMINING	33
4	Settings	34
4.1	BASIC SETTINGS.....	34
4.2	ADVANCED SETTINGS	35
4.3	CONNECTION	36
4.4	CHARACTER HANDLING.....	36
4.5	TERMINOLOGY HARVESTING	37
5	Uninstalling.....	39
6	Index.....	41

1 Introduction

In this chapter:

About crossMining (see below)

About this documentation (page 4)

1.1 About crossMining

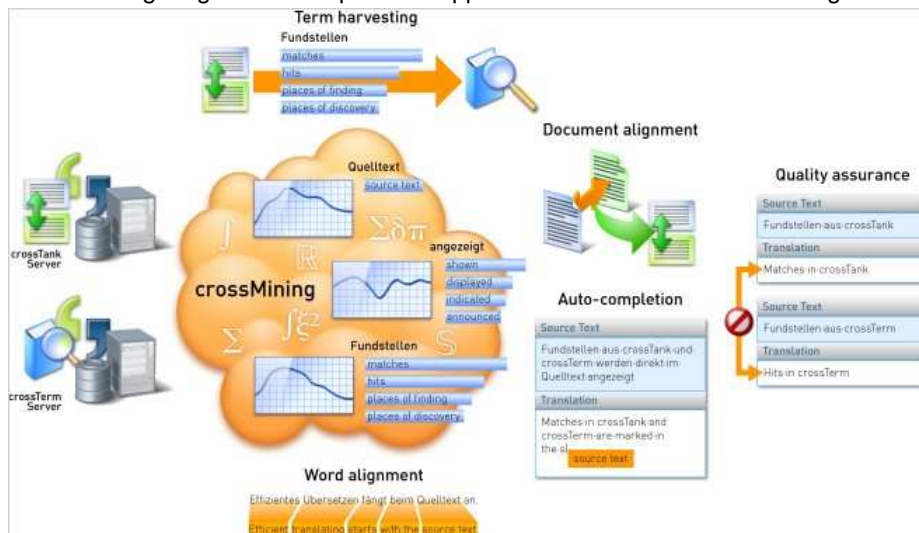


Statistical added value

crossMining is a tool that examines the linguistic resources of the Across Language Server and statistically analyzes the contents of crossTank entries for correlations between the source and target languages. For example, a probability calculation is used to automatically determine matching words in the translation units of a bilingual translation memory (e.g. English *engine* and German *motor*).

Bilingual statistical lexica containing the equivalents determined in the course of the lexicon creation represent one of the intermediate results of the work of crossMining. The statistical lexica can be used for various functions directly in crossMining and within Across. The application fields of crossMining range from the creation and/or supplementation of termbases to autocompletion while translating in crossDesk.

The following diagram shows possible application scenarios of crossMining:



Further information on crossMining and its various functions is available in this manual.

1.2 About this Documentation

This manual addresses users who want to work with crossMining.






This manual does not contain general information on the use of Across. For these instructions, please consult the Across user manuals and the Across Online Help. The latest version of the Across user manuals is available for download on the Across Web site at www.across.net/en/documentation-center.aspx!



This documentation was created using OfficeHelp.
www.officehelp.de

1.2.1 Icons

This manual makes use of icons and conventions to facilitate orientation.

Icon	Meaning
	Attention This icon indicates information that is essential for the correct use of crossMining.
	Tip This icon indicates tips and useful recommendations that facilitate the work with crossMining.
	Additional information This icon points to additional information and explanations intended to improve your understanding of the described feature.
	Cross-reference This icon points to more detailed information in other chapters or documents.
	New features and changes in Across v5.0 SP1 This icon marks new features and changes in crossMining version 5.0 Service Pack 1. Moreover, it points to extensions in the documentation (e.g., added chapters).

1.2.2 Conventions

For improved legibility and clarity, this manual makes use of the following spelling conventions:

- Key labels, names of menus and commands are presented in **bold and spaced** typeset.
- Technical terms are printed in *italics*.

1.2.3 Cross-reference

As Across and crossMining are subject to ongoing development, the documentation, too, is constantly being expanded and updated. For the latest version of the documentation and further Across-related information, go to www.across.net.

1.2.4 Feedback

Our objective is to provide all crossMining and Across users with optimum working conditions. For this reason, we always appreciate any feedback you send us. All information, texts, and illustrations have been prepared with utmost care. Nevertheless, errors may occur. If necessary, please contact documentation@across.net.

1.2.5 Document Versions

crossMining version	Document version	date	Changes
1.0	1.0	2009-09-03	Manual creation
1.0.5e	1.1	2009-10-30	Content update and extension
1.1.13	2.0	2010-07-12	Content update and extension

2 Installation

In this chapter:

System requirements (see below)
Installing crossAPI Interactive (page 6)
Installing crossMining (page 12)

2.1 System Requirements

As especially the creation of statistical lexica (see page 17) is a resource-intensive process, the computer on which crossMining is to be installed on the server side should be equipped accordingly. The system requirements are similar to those for an Across Server. (The latest version of the system requirements is available at www.across.net/en/documentation-center.aspx.)

Also, the installation of Microsoft .NET Framework 3.5 SP1 is required for using crossMining. (It is installed by default.)



crossMining can be installed and used directly on the server, i.e. on the computer that also hosts the Across Language Server, or on another computer or client in the local network. The communication between the Across Language Server and crossMining is handled by crossAPI Interactive, the open interface for real-time access to crossTank and crossTerm. For this reason, crossAPI Interactive must be installed before installing crossMining.



Follow the instructions starting on page 6.

2.2 Installing crossAPI Interactive

The following instructions describe the installation of crossAPI Interactive, the interface of Across for real-time access to crossTank and crossTerm.

Before installing crossAPI Interactive, make sure it is not already installed on your computer. To do this, check whether the entry "crossAPI Interactive" exists under "Programs and Features".



For further information on how crossAPI Interactive can be used please contact our support at www.across.net/en/form-supportrequest.aspx.



To install crossAPI Interactive, you need a generic softkey, which you will be required to enter in the course of the installation. This softkey is responsible for authenticating the crossAPI Interactive user at the Across Server.

If you do not have a generic softkey, please contact your Across system administrator, who will be able to create one for you.



If you are an Across system administrator, follow the instructions starting on page 9 to create a generic softkey.

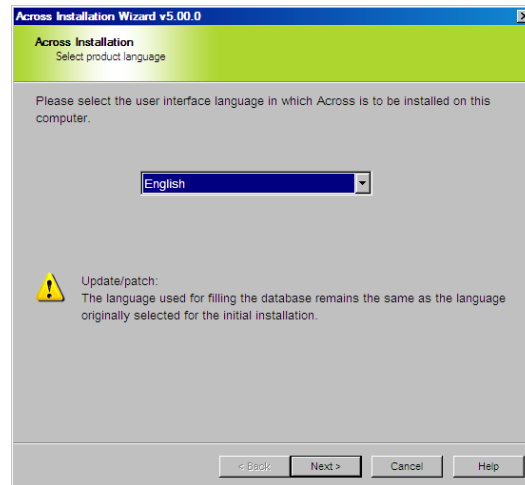
1. Log in to your PC as a user with administrator rights.
2. If necessary, unpack the file **Across_v50_en_crossMining.zip** (e.g. if you downloaded the installation files) and save the extracted files to your hard disk.

2 Installation

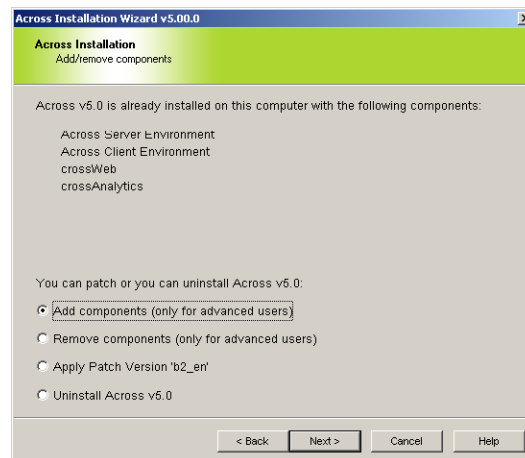
Installing crossAPI Interactive



3. Double-click the file **setup.exe** to launch the Installation Wizard that will lead you through the installation of crossAPI Interactive.
To install the software under Windows Vista, be sure to run the file **setup.exe** with administrator permissions. To do this, right-click the file and select the command **Run as administrator** from the context menu.
4. Once the wizard has started, click **Next>**.
5. If necessary, select the language in which you want to install crossAPI Interactive and click **Next >**.



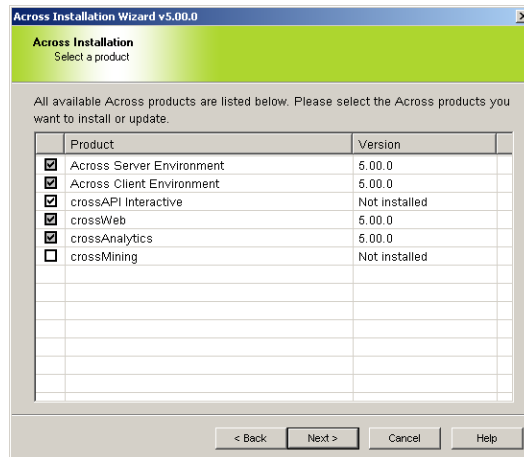
6. Enable the checkbox to confirm that you have read the information and wish to continue with the installation. Click **Next >** and select the language in which you want to use Across.
7. Mark the checkbox to confirm that you have read the license agreement (EULA) and accept it. Then click **Next >**.
8. Select the user-defined installation or the option for adding components and click **Next >**.



2 Installation

Installing crossAPI Interactive

9. Enable the corresponding checkbox to signify your wish to install crossAPI Interactive. Then click **Next >**.



10. Click **Next >** to continue installing crossAPI Interactive.



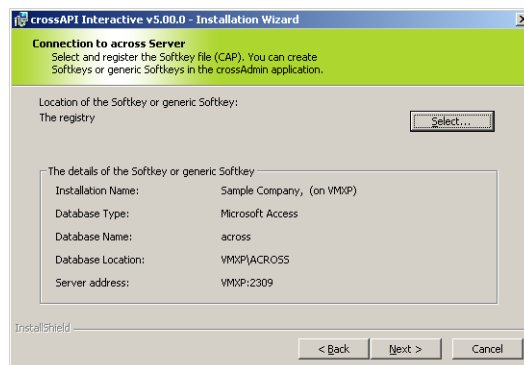
11. Now the generic softkey is defined. Usually, it is automatically generated and detected by Across. This softkey is responsible for authenticating the crossAPI Interactive user against the Across Server.



If the generic softkey was not automatically generated and detected, you must first create it and select it via **Select....**

Instructions on creating generic softkeys are available starting on page 9.

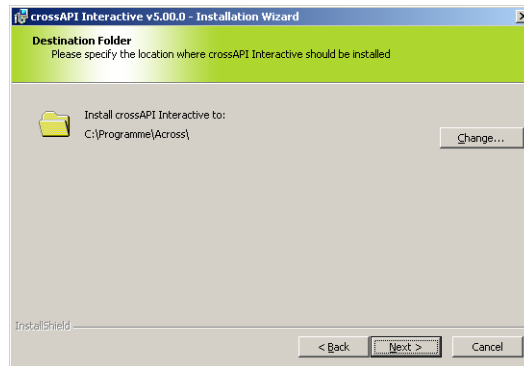
Click **Next >** to continue with the installation.



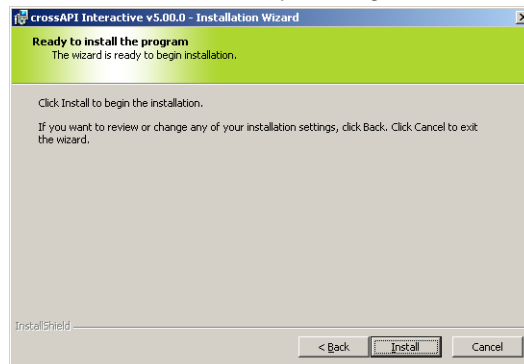
2 Installation

Installing crossAPI Interactive

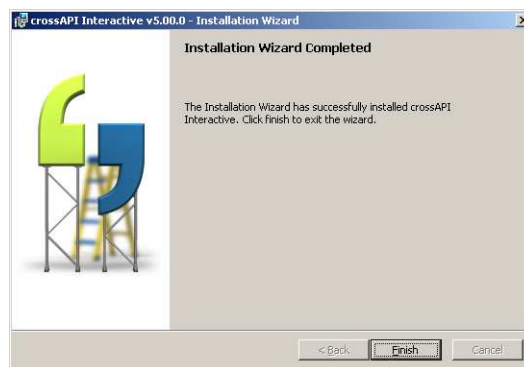
12. A location for the installation of crossAPI Interactive will now be suggested. To change the location, click **Change...** and select a different location. Then click **Next >**.



13. Launch the installation by clicking **Install**.



14. Click **Finish** to conclude the installation.



15. crossAPI Interactive has been successfully installed.



16. Now you can go to page 12 to continue with the installation of crossMining.

2.2.1 Creating a Generic Softkey

The generic softkey is created in crossAdmin – the administration software for the Across Server. You can save the generic softkey to a storage medium (e.g., hard disk) or send it by e-mail directly from crossAdmin.

2 Installation

Installing crossAPI Interactive

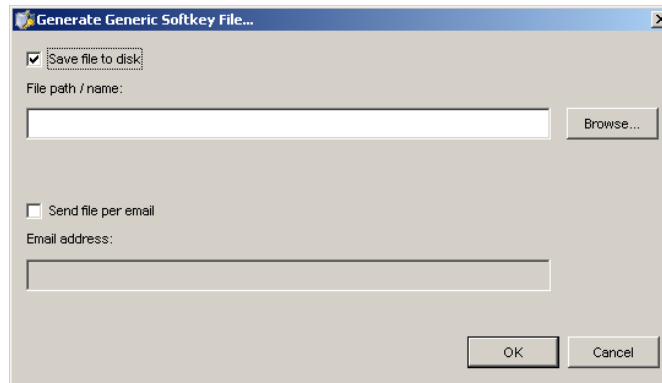


In most cases, only the Across system administrator has access to crossAdmin. Please contact the system administrator if you need a generic softkey.

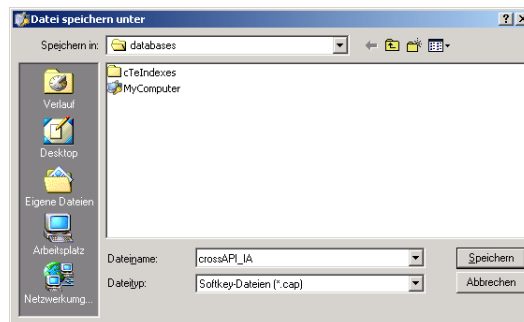
Follow these instructions to save the softkey to a storage medium.
Follow the instructions on page 11 to send the softkey by e-mail.

Saving the Generic Softkey to a Storage Medium

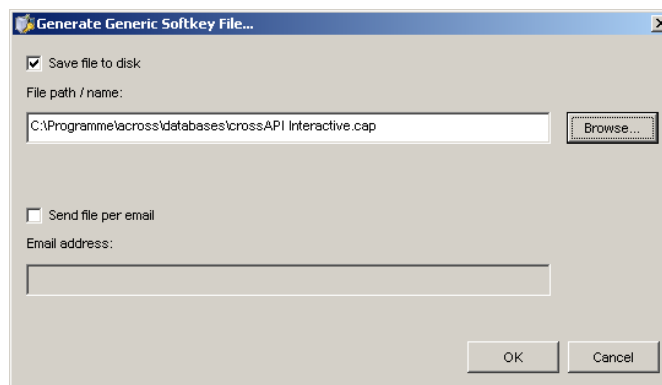
1. Open the crossAdmin administration application via **>>Start >>All Programs >>Across >>crossAdmin**.
2. Select the menu item **>>Tools >>Create generic softkey...**
3. Enable the option **Save file to disk** and then click **Browse...**



4. Select a location and enter a name for the softkey. Then click **Save**.



5. Click **OK**.



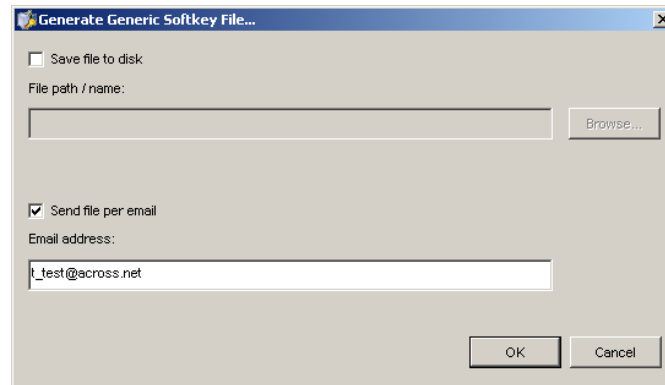
6. The generic softkey has now been created.



Sending the Generic Softkey by E-Mail

To be able to send generic softkeys by e-mail, the e-mail address of the Across Server and the SMTP server must be entered in crossAdmin under **>>Tools >>Settings... >>E-mail!**

1. Open the crossAdmin administration application via **>>Start >>All Programs >>Across >>crossAdmin.**
2. Select the menu item **>>Tools >>Create generic softkey...**
3. Enable the option **Send file by e-mail.** Then enter the e-mail address to which the softkey should be sent and click **OK.** The softkey will then be sent.



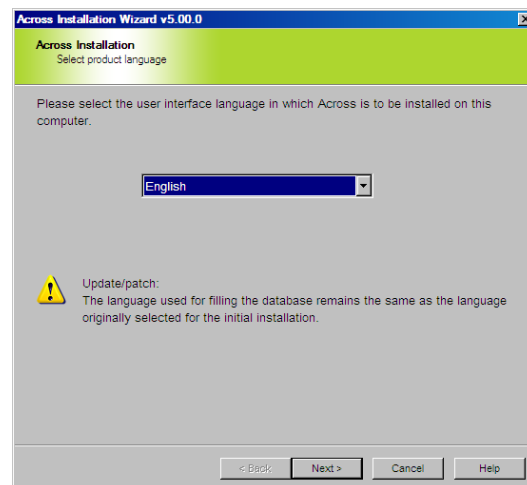
2.3 Installing crossMining

The following instructions describe the server-side installation of crossMining.

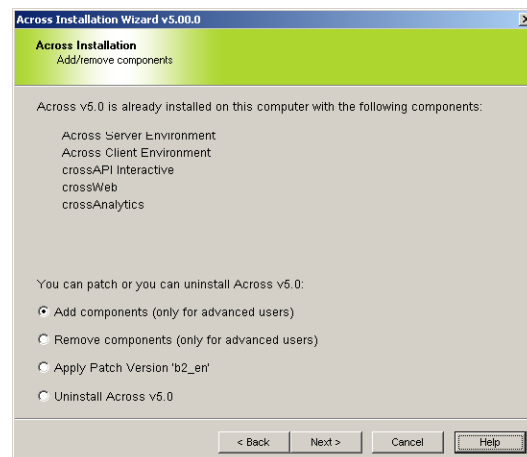
1. Log in to your PC as a user with administrator rights.
2. If necessary, unpack the file **Across_v50_en_crossMining.zip** (e.g. if you downloaded the Across installation files) and save the extracted files to your hard disk.
3. Double-click the file **setup.exe** to launch the Installation Wizard, which will guide you through the installation of crossMining.

To install the software under Windows Vista, be sure to run the file **setup.exe** with administrator permissions. To do this, right-click the file and select the command **Run as administrator** from the context menu.

4. Once the wizard has started, click **Next >**.
5. If necessary, select the language in which you want to install crossMining and click **Next >**.



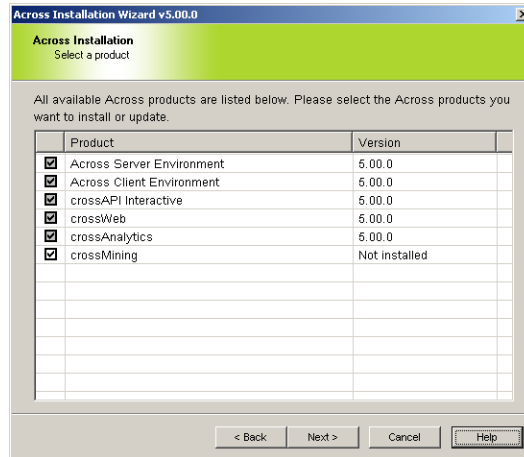
6. Enable the checkbox to confirm that you have read the information and wish to continue with the installation. Then click **Next >**.
7. Mark the checkbox to confirm that you have read the license agreement (EULA) and accept it. Then click **Next >**.
8. Select the user-defined installation or the option for adding components and click **Next >**.



9. Mark the checkbox for installing crossMining. Then click **Next >**.

2 Installation

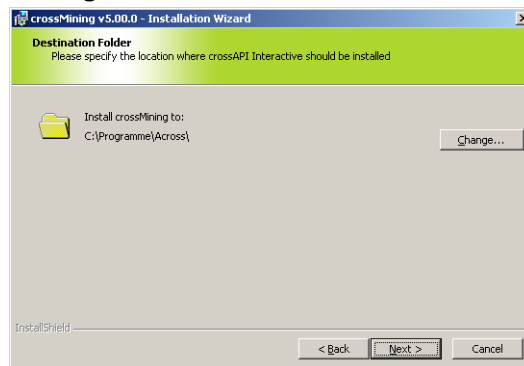
Installing crossMining



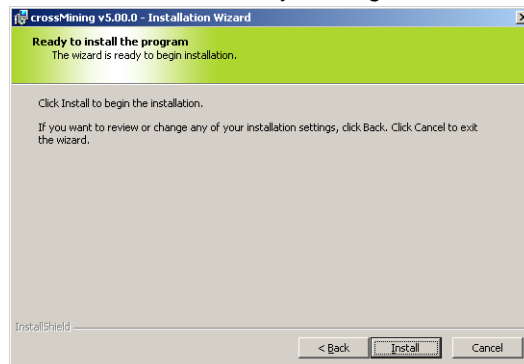
10. Click **Next >** to continue with the installation of crossMining.



11. A location is proposed for installing crossMining. To change the location, click **Change...** and select a different location. Then click **Next >**.



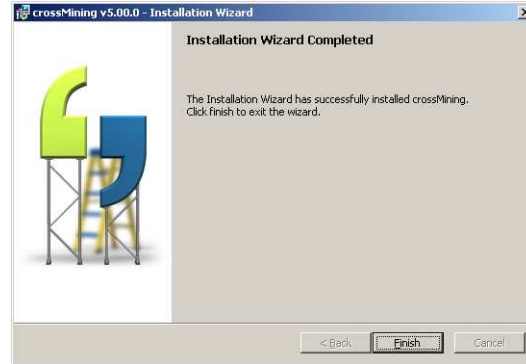
12. Launch the installation by clicking **Install**.



2 Installation

Installing crossMining

13. Click **Finish** to conclude the installation.



14. crossMining has been successfully installed.



3 Using crossMining

In this chapter:

Overview of crossMining functions (see below)

Starting crossMining (page 15)

Working with crossMining (page 17)

Closing crossMining (page 33)

Overview of crossMining

3.1 Overview of crossMining Functions

Statistical lexica (see page 17) form the basis for the work with crossMining. These can be created with crossMining and used for other functions in crossMining and within Across. Terminology harvesting (see page 24) for the semi-automatic expansion of the terminology base is one of the application fields directly in crossMining. The bilingual term extraction (see page 27) enables the creation of entirely new terminology entries. Moreover, existing terminology bases can be expanded with additional target-language terms with the help of the addition of target-language terms (see page 24).

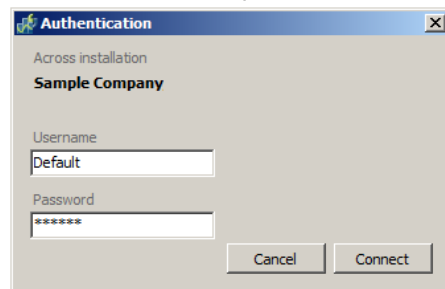
However, the statistical lexica created with crossMining can also be used directly for translating with Across: Thanks to the autocompletion function (see page 21), the contents of the statistical lexica are proposed to the translator while working in crossDesk, allowing the translator to benefit directly from the crossMining results.

3.2 Starting crossMining

Proceed as follows to start crossMining:

1. Start crossMining via **>>Start >>All Programs >>Across >>crossMining**.
2. Specify the username and, if necessary, the password of the user over whom the Across Server is to be accessed via crossMining. This user must be member of the group "crossAPI Interactive read/write access" in Across.)

crossMining automatically uses the Across Server selected by means of the generic softkey during the installation of crossAPI as the Across Server whose data are to be accessed. (To connect to another server, click **Cancel** and select the desired server in the connection settings under **>>Tools >>Settings... >>Connection**. Further information on this topic is available on page 36.)



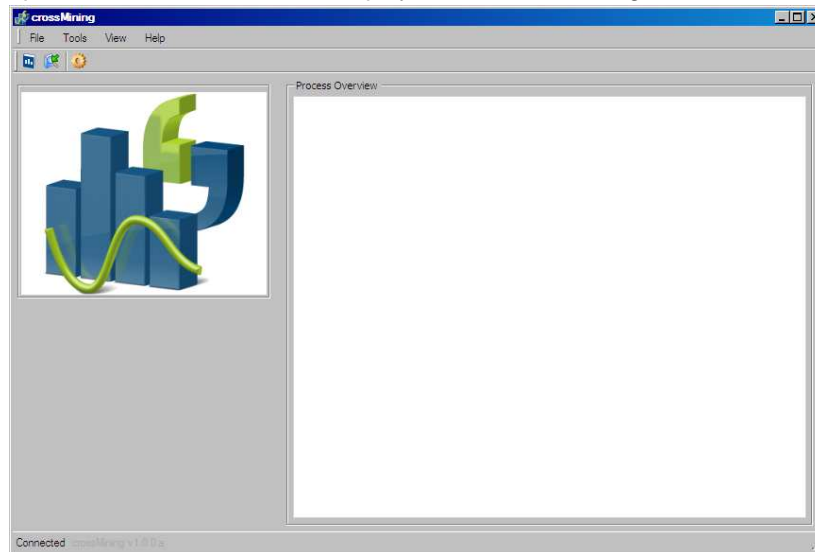
3. Click **Connect** to connect crossMining to the selected Across Server.

Windows Firewall

If you want to install crossMining on a computer protected by a Windows firewall, a firewall alert will appear after crossMining is started for the first time. Please confirm the alert with **Unblock**.

Please note that in the Microsoft firewall settings under **>>Start >>Settings >>Network connections >>LAN connection >>General >>Properties >>Advanced >>Settings >>Exceptions**, the option "Display a notification when Windows Firewall blocks a program" must be enabled. If this option is disabled, please enable it before installing Across. If you are not sure how to proceed, please contact your system or network administrator.

4. Following the establishment of the connection to the Across Server, crossMining opens up. The "Connected" state is displayed at the bottom edge of the screen.






3.3 Working with crossMining

In this chapter:

crossMining toolbar (see below)
 Statistical lexica (page 17)
 Autocompletion (page 21)
 Terminology harvesting (page 24)
 Import of the Moses SMT phrase tables (page 30)

3.3.1 crossMining Toolbar

The crossMining toolbar offers the following functionalities:

Icon	Meaning
	Creating statistical lexica (see page 18)
	Starting terminology harvesting (see page 24)
	Opening crossMining settings (see page 34)

3.3.2 Statistical Lexica

Language is what counts...



Statistical lexica form the basis for the work with the various functions of crossMining. These are created automatically in several steps and are mainly based on the crossTank data of an Across Language Server. Optionally, the existing terminology in crossTerm can also be taken into consideration when creating lexica.

Furthermore, statistical lexica can be created on the basis of Moses SMT phrase tables, a free system for statistical machine translation (see page 30).

The statistical lexica have the file extension ***.dic** and are created for a particular language pair. The lexica can only be used in one direction for the other crossMining functions, i.e. only for the language direction selected during creation.



Before you continue using the statistical lexica for the other functions of crossMining, you should test the lexicon creation thoroughly on the basis of your specific data and, if necessary, with professional help in order to ensure the most suitable values and settings for your data. The Across Professional Services team, which you can contact by e-mail to professional-services@across.net, will be pleased to assist you in this regard.

A certain amount of data (translation units) is necessary for the efficient, quality use of crossMining. The smaller the amount of data available for the calculation of probabilities, the poorer the results will be. Generally, about 10,000 translation units (per language pair) should be provided, though this does not mean that good results cannot be achieved with fewer translation units.

The quality of the results also depends on the respective language or language combination. Languages with a simpler morphological structure, such as English, enable good results even with a relatively small amount of data. In contrast, the satisfactory determination of

probabilities for highly inflectional languages like Finnish is only possible from a larger amount of training data. Moreover, the language direction is also important.



As the creation of the lexicon is very resource-intensive, it may take some time, depending on the data volume. Therefore, you should only run the lexicon creation at times when the computer has nothing or little else to do.




Of course, it is possible to create statistical lexica as often as necessary. Creating new lexica is recommended especially when the crossTank data have changed substantially, e.g. after importing a large translation memory or upon completion of a major translation project. Some users may want to create lexica at regular intervals, e.g. once a month.

Creating Statistical Lexica

Lexicon creation

Proceed as follows to create a statistical lexicon:

1. Start the lexicon creation via the  icon in the crossMining toolbar or via the menu item **>>File >>Create Lexicon...**

When creating lexica, the settings defined under **>>Tools >>Settings...** are used.

Further information on this topic is available in the respective chapter on page 34.

2. First, the basic settings are defined for the lexicon to be created.



Defining languages

The first step is the selection of the languages in which the lexicon is to be created.

crossMining automatically determines the languages set up in Across. Select a source language and then a target languages (and sublanguages if applicable).

You can freely combine the source and target languages and also define multiple language pairs. A separate lexicon is created for each language pair.



Setting filters

Now you can define crossTank and/or crossTerm filters to limit the lexicon creation to certain crossTank and crossTerm entries or ranges.

For crossTerm, you can filter the crossTerm data by instances, relations, and subjects.

For crossTank, you can filter by users, subjects, projects, relations, and user-defined system attributes.

Output directory

Subsequently, you can set the output directory for the lexicon. By default, a subdirectory of the "Common Files" directory in the "Program Files" folder is used for this purpose.

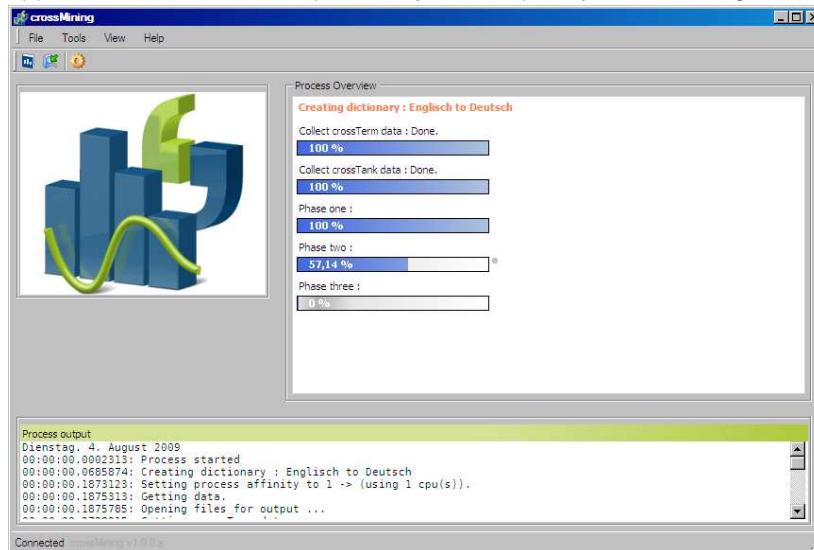


From this subdirectory, the statistical lexica are read and deployed to the Across Clients via autopatching. (Further information on this topic is available on page 21.)

If you wish, you can select a different output directory. For example, this enables you to optimize the creation of the statistical lexica for test purposes before you store the lexica in the default output folder for deployment to the clients. To select a different output folder, disable the option "Use default output folder" and click **Browse...** to select a different folder.

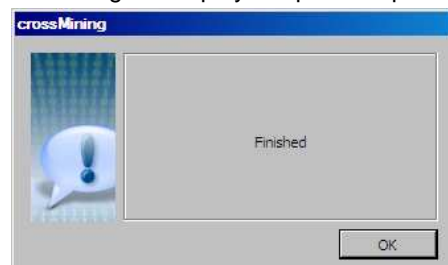
Then click **Start** to start the creation of a lexicon.

3. Now the lexica are created. This process comprises the following steps:
 - 1) Compilation of the crossTerm data in the selected languages (this step is skipped if the option for including terms (see above) is disabled).
 - 2) Compilation of the crossTank data in the selected languages.
 - 3) First phase of the lexicon creation: probability calculation of possible word equivalents.
 - 4) Second phase of the lexicon creation: inclusion of the word position in the probability calculation.
 - 5) Third phase of the lexicon creation: Determination of possible equivalents of multiple-word combinations (e.g. English *table of contents* vs. German *Inhaltsverzeichnis*) under application of the minimum probability and frequency values configured in the settings.



Under **>>View >>Process output**, you can have the process steps currently performed by crossMining displayed in a pane.

4. A message is displayed upon completion of the lexicon creation. Click **OK**.



5. The statistical lexicon has been saved to the selected storage location as .dic file. The name of the file consists of the installation GUID of the Across Language Server and the country codes (LCIDs) of the source and target languages.

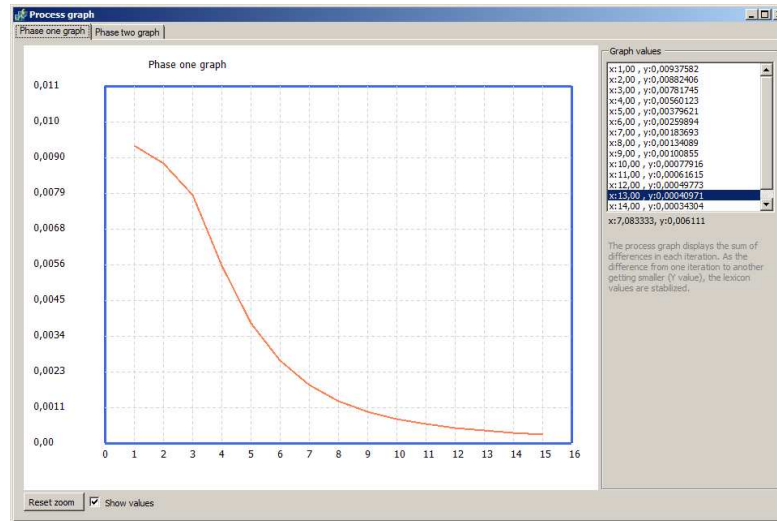
Now you can use the created lexicon for the autocompletion function (see page 21) and for terminology harvesting (see page 24) in crossMining.

3 Using crossMining

Working with crossMining

Process Graphs

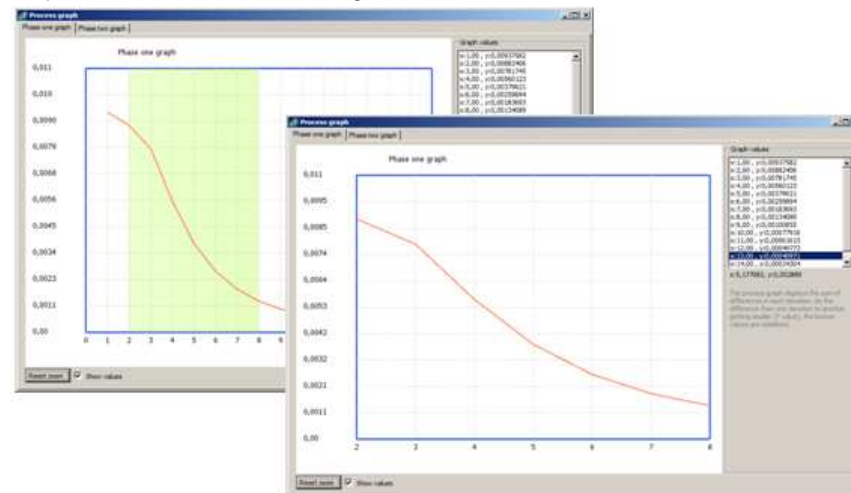
Under **>>View >>Graph**, you can view the development of probabilities during the generation of lexica in graphical form. The tabs allow you to select the graph for the first or second phase of the lexicon creation. The iterations are displayed on the x axis and the probabilities on the y axis.



The creation of statistical lexica can be optimized by analyzing the graphs, e.g. by duly adjusting the number of iterations.

Selecting Sections

If you select a section of the graph while keeping the left mouse button pressed, the respective section will be enlarged.



Click **Reset zoom** to restore the original display.

Deploying lexica**Deployment of Statistical Lexica**

Statistical lexica stored in the default output folder of the Across Server are automatically recognized by Across and deployed to the Across Clients via autopatching. Subsequently, the lexica can be used for the autocompletion function while translating in the Target Editor of crossDesk (see following chapter).

Online Clients & crossWAN load

When a client connects to the Across Server, the date and time of the statistical lexica are automatically compared with those on the server. If the lexica of the client are older than those of the server, they will automatically be transferred and stored in the corresponding folder.

**Lexicon folder**

The default output folder in which the lexica must be stored on the server side for automatic deployment to the clients is **Program Files/Common Files/Across/crossMining/dic**.

To use the autocompletion function on the client side, the files are stored or must be stored manually in the identical directory (see below).

Deployment for crossWAN classic...**Deployment of Lexica for crossWAN classic and for crossGrid**

When using crossWAN classic, the statistical lexica cannot be transmitted to the clients via autopatching, as the clients are not connected directly to the Across Server. Therefore, the lexica must be sent to the user in a different way (e.g. by e-mail) and then stored manually in the corresponding folder (see above).

... and for crossGrid

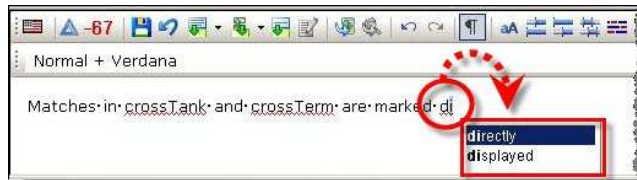
When using crossGrid (online and classic), the lexica must also first be transmitted manually (e.g. by e-mail) from the Master Server to the Trusted Server and stored in the corresponding folder of the Trusted Server, as crossGrid servers are not autopatched. From the Trusted Server, the lexica can be further distributed via autopatching (see above).

3.3.3 Autocompletion

The statistical lexica created with crossMining are deployed to the clients via autopatching (see page 21) and can subsequently be used while translating in crossDesk.

crossMining in Across**Autocompletion Function in Across**

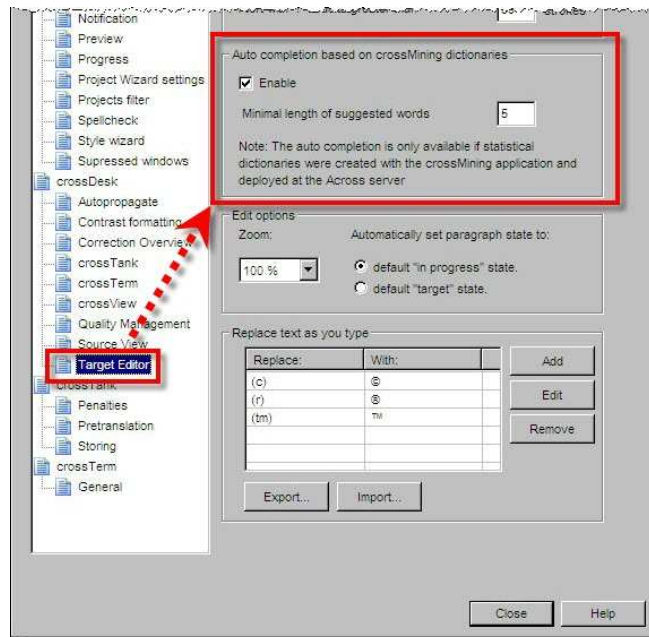
During the translation in crossDesk, the system supplies the translator with proposals that can quickly and easily be transferred to the target text via autocompletion. The proposals come from the lexica created with crossMining. The proposals may consist of individual words or entire sentence segments.



In Across, the autocompletion function can be enabled and configured in a subsection of the profile settings of Target Editor. There, you can determine that only words with more than a certain number of characters are to be proposed in the Target Editor during the translation. In this way, you can prevent short words like articles and prepositions from being proposed.

3 Using crossMining

Working with crossMining



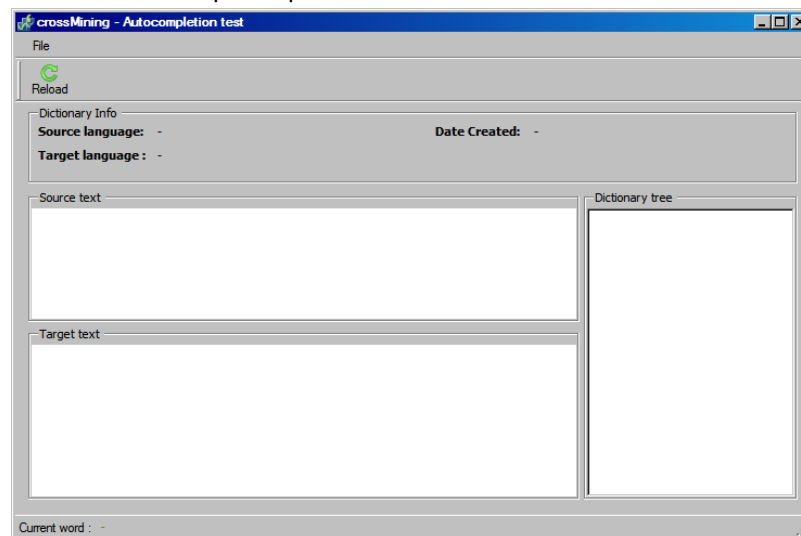
By default, the lexica are stored in the directory **Program Files/Common Files/across/crossMining/dic** on the Across Client side.

Autocompletion Test

crossMining features an integrated test function for checking the functionality and quality of the autocompletion function on the basis of the created statistical lexica.

Proceed as follows to employ the test mode:

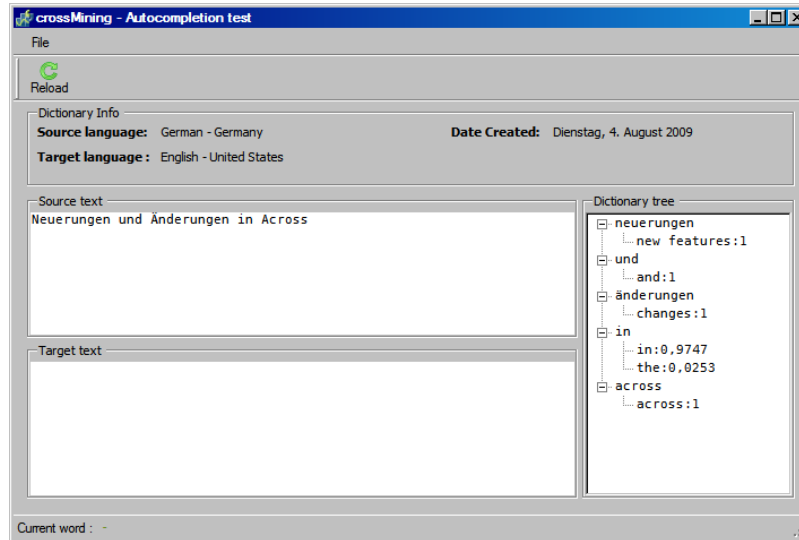
1. Start the test function via the menu item **>>File >>Autocompletion Test...**
2. The test window opens up.



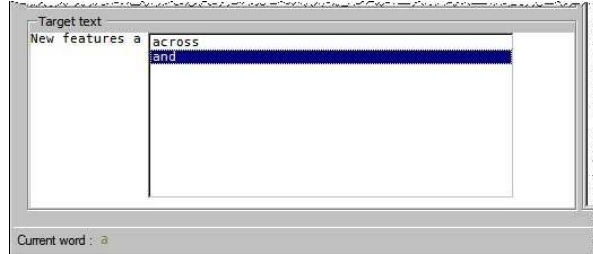
3. Go to **>>File >>Load lexicon** to select a lexicon for the test.
4. In the top left pane, insert a source text with which you want to test the autocompletion and click **Reload**.
The equivalents found will be displayed in the right window along with their probability of correspondence.

3 Using crossMining

Working with crossMining



5. Enter a translation of the source text in the bottom left pane. crossMining proposes any potential equivalents in a drop-down list.



6. Select a suitable equivalent with the arrow keys and transfer it to the translation by clicking **Enter**.



Harvesting terminology

3.3.4 Terminology Harvesting

Using the terminology harvesting functions, the terminology in crossTerm can be expanded in two different ways on the basis of the previously created statistical lexica (see page 17): Using the addition of target-language terms (see following chapter), existing terminology bases can be expanded with target-language equivalents in an additional language. The most probable target-language equivalents are proposed in crossMining and can subsequently be created automatically as terms in crossTerm.

Moreover, the bilingual term extraction (see page 27) can be used to create entirely new entries. In this context, source-language term candidates are proposed with the probable target-language equivalents in crossMining and can subsequently be created automatically as terms in Across.



The terminology harvesting settings can be configured in the new "Terminology harvesting" section of the crossMining settings (under **>>Tools >>Settings...**). Further information is available in the corresponding chapter on page 37.

Adding Target Terms


Addition of Target-Language Terms

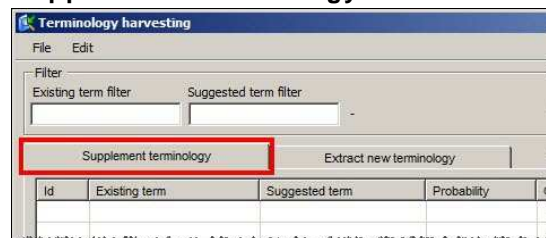
The addition of target-language terms enables the expansion of existing termbases with target-language equivalents in an additional language. If, for example, the corresponding English terms are to be added to an existing German termbase, crossMining will automatically extract possible English translations of the existing German terms and propose these to the user.

After a target-language term candidate is automatically extracted and proposed, the user can confirm it as a term. Upon confirmation, the new term is automatically sent to crossTerm, where it is created as a term.

When transmitted from crossMining, the terms added in this way are set to "unreleased". A user currently logged in to crossMining will be registered as the creator of the terms. Thus, the new terms can subsequently be searched for systematically (e.g. by using the filter for searching for unreleased terms and/or for terms created by a respective user), edited, and released.

Proceed as follows to add target-language terms:

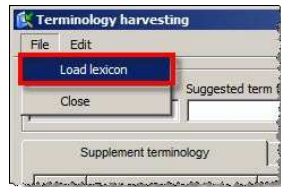
1. Launch the addition of target-language terms by clicking the  icon in the crossMining toolbar or via the menu item **>>File >>Terminology Harvesting...**
2. The terminology harvesting dialog appears. Target-language terms can be added in the **Supplement terminology** tab.



3 Using crossMining

Working with crossMining

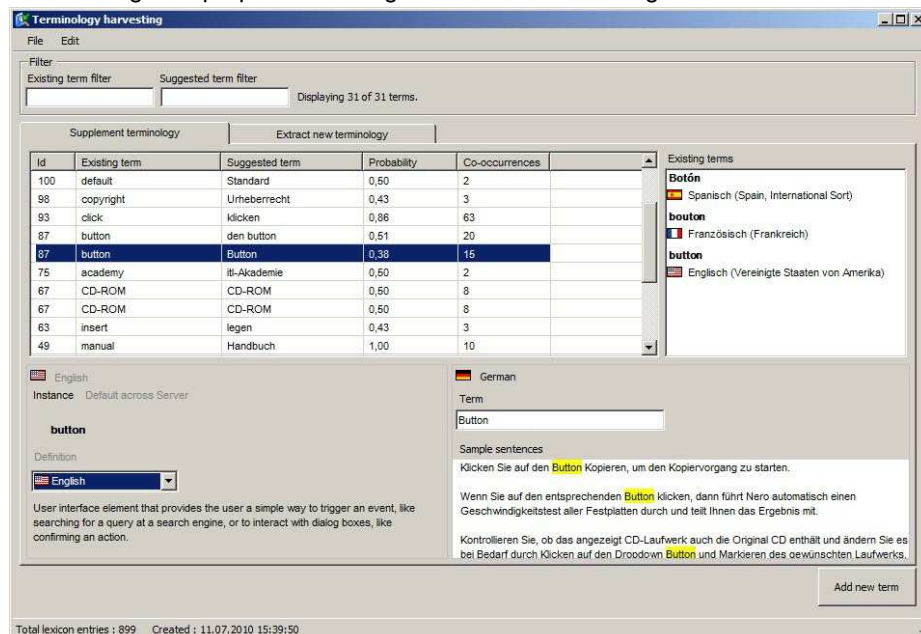
3. First, select the statistical lexicon that you want to use the basis for the addition of target-language terms. To do this, click **>>File >>Load lexicon**.



4. A dialog window lists all statistical lexica stored in the output directory. (Click **Select path** to select a different directory.)
Select the lexicon you would like to use.
Furthermore, you can select the crossTerm instances to be taken into consideration when adding terms. Target-language terms will only be added for source-language terms contained in the selected instances.
Finally, you can determine the minimum frequency and probability from which terms are to be proposed and stopwords not to be accounted for.
Click **OK** to start the addition of target-language terms.



5. crossMining now proposes the target terms for the existing source terms.





In addition to the proposed target terms and the existing source terms, the probability of correspondence between the source and target-language terms, the co-occurrence count, and the IDs of the respective entries in crossTerm are displayed.

Click one of the column headers of the table to change the table sorting on the basis of the selected column.

You can narrow down the list of displayed source and target terms by entering one or several characters in the filter input fields. Only the source/target terms beginning with these letters will be displayed. To limit the list to terms ending in particular characters, you can use the asterisk (*) (e.g. ***ion** to display only terms ending in "ion"). To limit the list to terms containing one or several characters, you can place the asterisk at the beginning and end of the filter string (e.g. ***r*** to display only terms containing the letter "r").

Filter
Existing term filter: C Suggested term filter: Displaying 8 of 31 terms.

Supplement terminology | Extract new terminology

Id	Existing term	Suggested term	Probability	Co-occurrences
6	CD-ROM	CD-ROM	0,50	8
93	click	klicken	0,86	63
96	contact	unter	0,40	2
98	copyright	Urheberrecht	0,43	3

Moreover, the context in which the target terms are used in crossTank entries and the terms in other languages that already exist in crossTerm are displayed.

Sample sentences

Informationen zu **Urheberrecht** und Marken

Dieses Handbuch enthält Materialien, die durch international geltendes **Urheberrecht** geschützt sind.

Any available definition for the source term or entry is displayed. If there are several definitions, you can select them from the drop-down list.

Existing terms

copyright
 Englisch (Vereinigte Staaten von Amerika)

copyright
 Französisch (Frankreich)

Derecho de propiedad intelectual
 Spanisch (Spain, International Sort)

6. Select the target term(s) (or the respective table rows) you want to add to the respective entries in crossTerm. (Use the **Ctrl** or **Shift** key for multiple term selection.)

If necessary, you can manually correct the proposed terms.

7. Click **Add new term** and confirm the subsequent message with **Yes**.



The target term(s) is/are sent to crossTerm and added to the source-language term(s). Every term is assigned the picklist values and text fields defined in the terminology harvesting settings (under **>>Tools >>Settings... >>Terminology harvesting**). Furthermore, the terms are set to "unreleased". The user currently logged in to crossMining will be entered as the author of the terms.

- A message confirms the successful creation of the term. Click **OK**.



- The terms that were just created in crossTerm are removed from the list. Continue until you have added all desired target terms to crossTerm. Click **>>File >>Close** to finish the addition of target-language terms.




Extracting terms – in two languages!

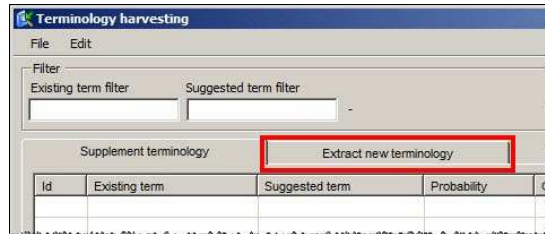
Bilingual Term Extraction

In addition to the normal monolingual term extraction within Across, crossMining enables an additional term extraction in which the source-language term candidates are already proposed with their potential target-language equivalents.

After a term candidate pair is automatically extracted and proposed, the user can confirm it as a term pair. Upon confirmation, the term pair is automatically sent to crossTerm, where it is created as a new terminology entry. The terms are set to "unreleased". A user currently logged in to crossMining will be registered as the creator of the entries and terms. Thus, they can subsequently be searched for systematically (e.g. by using the filter for searching for unreleased terms and/or for terms created by a respective user), edited, and released.

Proceed as follows to perform a bilingual term extraction:

- Start the bilingual term extraction via the  icon in the crossMining toolbar or via the menu item **>>File >>Terminology Harvesting...**
- The terminology harvesting dialog appears. The bilingual term extraction can be performed in the **Extract new terminology** tab.



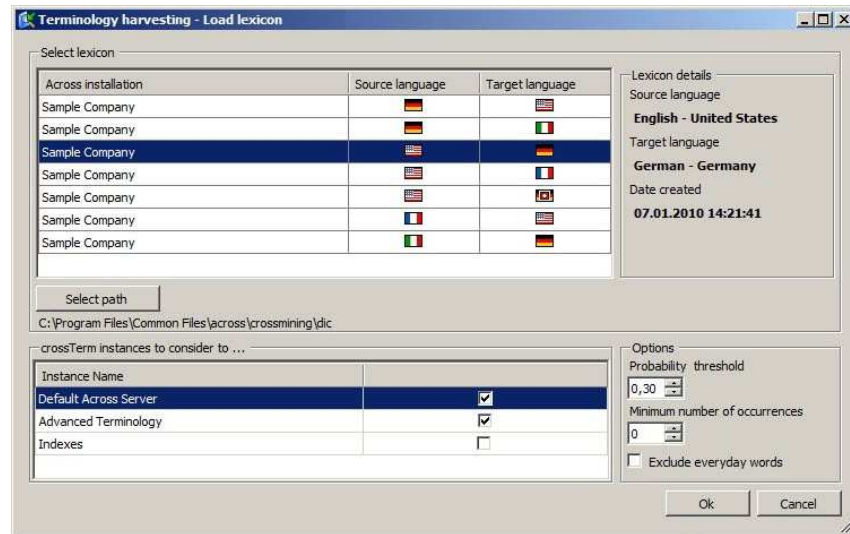
- First, select the statistical lexicon that you want to use as the basis for the term extraction. To do this, click **>>File >>Load lexicon**.



- A dialog window lists all statistical lexica stored in the output directory. (Click **Select path** to select a different directory.)
Select the lexicon you would like to use.
Finally, you can determine the minimum frequency and probability from which terms are to be proposed and stopwords not to be accounted for.
Click **OK** to start the term extraction.

3 Using crossMining

Working with crossMining



5. crossMining now proposes term pairs.



In addition to the proposed source and target-language terms, the probability of correspondence between the source and target-language terms and the co-occurrence count are displayed.

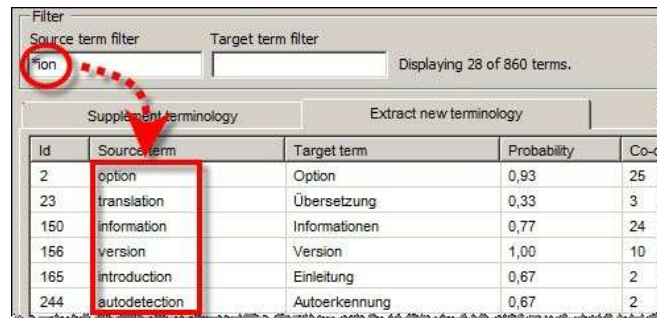
Click one of the column headers of the table to change the table sorting on the basis of the selected column.

You can narrow down the list of displayed source and target terms by entering one or several characters in the filter input fields. Only the source/target terms beginning with these letters will be displayed. To limit the list to terms ending in particular characters, you can use the asterisk (*) (e.g. *ion to display only terms ending in "ion"). To limit the list to terms containing one or several characters, you can place the asterisk at the beginning and end of the filter string (e.g. *r* to display only terms containing the letter "r").



3 Using crossMining

Working with crossMining



The screenshot shows the crossMining interface. At the top, there is a 'Filter' section with 'Source term filter' and 'Target term filter' text boxes. The 'Source term filter' contains the text 'tion'. Below this is a table with columns: 'id', 'Source term', 'Target term', 'Probability', and 'Co-oc'. The table contains the following data:

id	Source term	Target term	Probability	Co-oc
2	option	Option	0,93	25
23	translation	Übersetzung	0,33	3
150	information	Informationen	0,77	24
156	version	Version	1,00	10
165	introduction	Einleitung	0,67	2
244	autodetection	Autoerkennung	0,67	2

Moreover, the context in which the source and target-language terms are used in crossTank entries are displayed.



The screenshot shows 'Sample sentences' with the following text:

Informationen zu **Urheberrecht** und Marken

Dieses Handbuch enthält Materialien, die durch international geltendes **Urheberrecht** geschützt sind.

6. Select the term pair(s) (or the respective table rows) you want to add in crossTerm. (Use the **Ctrl** or **Shift** key for multiple term selection.)

If necessary, you can manually correct the proposed terms.

7. Click **Add new entry** and confirm the subsequent message with **Yes**.

The term pair(s) is/are sent to crossTerm and created as new entries. The entries are created in the Across instance determined for this purpose in the terminology harvesting settings (under **>>Tools >>Settings... >>Terminology harvesting**).

Every term is assigned the picklist values and text fields that are also defined in the terminology harvesting settings. Furthermore, the terms are set to "unreleased". The user currently logged in to crossMining will be entered as the author of the terms.

8. A message confirms the successful creation of the entries. Click **OK**.



9. The terms that were just created in crossTerm are removed from the list. Continue until you have added all desired term pairs to crossTerm.

Click **>>File >>Close** to finish the bilingual term extraction.





3.3.5 Import of Moses SMT Phrase Tables

Service Pack 1 of Across now enables the import of phrase tables of Moses SMT, a free system for statistical machine translation.

On the basis of the phrase tables, statistical lexica can be created and used for terminology harvesting or autocompletion in crossDesk, just like the conventional lexica created on the basis of the crossTank data.



The phrase tables created with Moses SMT are text files containing source-language phrases (e.g. individual words, several words, or sentences) and their – statistically determined – target-language equivalents including statistical information.

The Dictionary Import Wizard assists you in creating a statistical lexicon on the basis of a Moses SMT phrase table.

Proceed as follows to import a Moses SMT phrase table and create a statistical lexicon:

1. Start the Dictionary Import Wizard via >>**File** >>**Import....**



2. Once the wizard has started, click **Next >**.



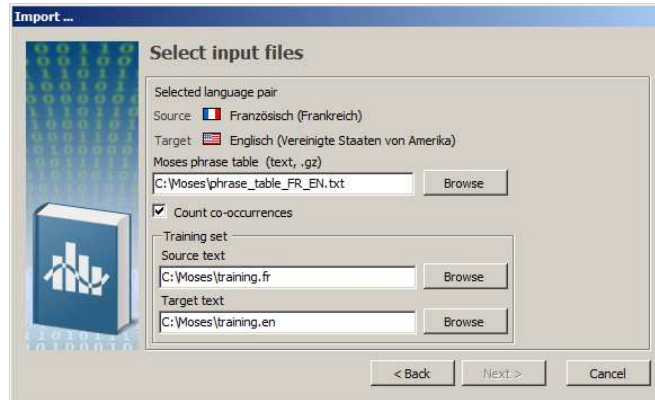
3. Select the source and target languages (and the sublanguage, if applicable) contained in the phrase table and click **Next >**.



4. Select the storage location of the phrase table by clicking **Browse...**

The phrase table may exist in the form of a plain TXT file or a compressed GZ file. Under the option "Count co-occurrences", you can determine a training set consisting of a parallel text pair (in the source and target languages). In the next wizard step, you can determine a minimum co-occurrence count – i.e. search hits both in the source text and in the target text of the training set – for the lexicon creation.

Click **Next >**.



5. You can now determine the minimum probability from which terms are to be proposed. Moreover, you can specify the minimum probability value of correspondence of the source and target-language terms.

Furthermore, you can determine that terms are to be proposed only above a specified co-occurrence count (see above).

Finally, you can exclude phrase-table entries from the lexicon creation. For this, you can define words that should not occur at the beginning or end of the respective entries in the source and target-language phrase-table entries. Click **Edit...** to determine the words. You can edit the words manually, import them from a file, and/or import the stopword list of the particular language from Across. Click **Save** to finish the definition of words.

Then click **Next >**.



6. Now you can set the output directory for the lexicon. By default, a subdirectory of the "Common Files" directory in the "Program Files" folder is used for this purpose.

Click **Start Import** to start creating the statistical lexicon.

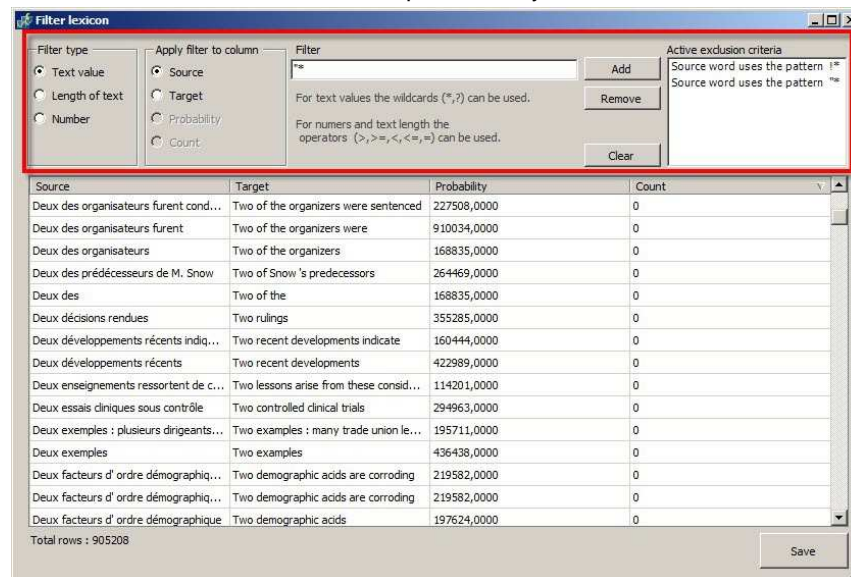
As the creation of the lexicon is very resource-intensive, it may take some time, depending on the size of the chosen phrase table. Therefore, you should only run the lexicon creation at times when the computer has nothing or little else to do.





7. Upon completion of the lexicon creation, the lexicon is displayed with the determined equivalents, the respective probability, and the co-occurrence count.

As Moses SMT phrase tables can be very large and contain several million entries, the statistical lexica generated on the basis of these can also be very large. Therefore, you can narrow down the determined equivalents by means of extensive filter functions.



8. To edit the created lexicon, you can define filter criteria.

First, select one of the following three filters:

- *Text value*: Filter on the basis of a particular text or character string.
- *Text length*: Filter on the basis of a particular number of characters.
- *Number*: Filter on the basis of the probability or co-occurrence count.

After selecting a filter, you can select the column to which the filter criterion is to be applied. For "Text value" and "Text length", you can select either the source text or the target text. For "Number", you can determine that the filter is to refer to the probability or to the co-occurrence count.

Subsequently, you can enter the respective value for the filter – e.g. a word or special characters (for "Text value") or a particular numeric value (for "Text length" and "Number"). In the latter case, you can use one of the following operators: > (greater than), >= (greater than or equal to), < (less than), <= (less than or equal to), = (equal to).

Click **Add** to adopt the filter criterion.

Select the filter type, ...

... column, and ...

... value



Please note that the filter process will take place immediately after adding a filter criterion. For large lexica, this might take some time.

9. Click **Save** to save the statistical lexicon to the selected output directory.
10. A message is displayed after the lexicon is saved to the output directory.



You can now use the lexicon for the autocompletion functions (see page 21) and the terminology harvesting functions (see page 24) of crossMining just like conventional lexica created on the basis of the crossTank data.

Further information on statistical lexica is available on page 17.



3.4 Closing crossMining

To close crossMining, click the menu item **>>File >>Exit** and confirm with **Yes**.



4 Settings

The function of crossMining as well as the access to crossMining can be configured in the crossMining settings. The settings comprise the following three tabs:

- Basic settings (see below)
- Advanced settings (see page 35)
- Connection (see page 36)
- Character handling (see 36)
- Terminology harvesting (see page 37)



You can access the crossMining settings via the  icon in the crossMining toolbar or via the menu item >>**File** >>**Settings**....

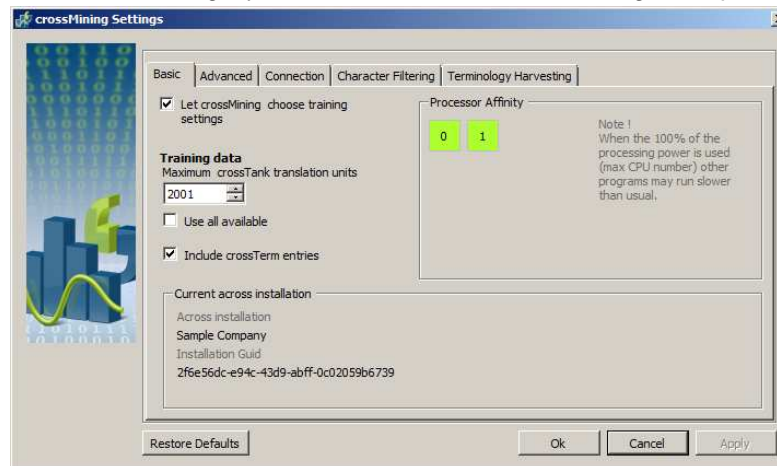
Click **Reset to default** to restore the original basic and advanced settings.

4.1 Basic Settings

In the basic settings, you can determine how crossMining is to operate when creating lexica.



Basic settings



First, you can determine that crossMining select the parameters for the generation of statistical lexica. In this case, crossMining will select the optimum number of iterations for phase 1 and 2 of the lexicon creation. (If this option is activated, the section for setting the iterations in the **Advanced** tab will be disabled – see page 35.)

Training data

Moreover, you can determine the maximum size of the training data, i.e. the maximum number of translation units/crossTank entries. Enter the desired number in the input field or use the arrow icons. If there are more translation units that the specified maximum, the translation units will be used in ascending chronological order, starting with the oldest translation units.

If you enable the respective option, all available translation units will be considered by the calculation.

Including crossTerm

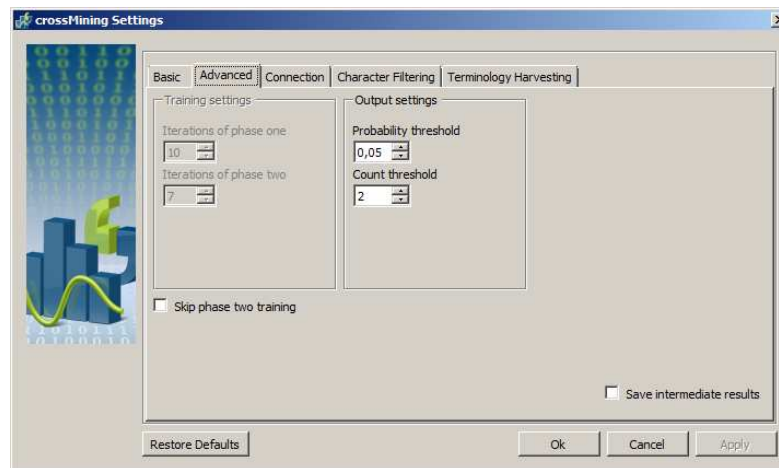
You can also determine whether the terms that already exist in the respective languages in crossTerm are to be included when creating the lexicon. In this case, all terms will be used as additional crossTank units and included in the probability calculation. Including the terms can deliver better results, especially when using the autocompletion function.

Number of CPU Cores

If your computer has multiple CPU cores, you can select the cores to be used when working with crossMining by clicking the desired core in the "Processor Affinity" area. When selecting CPU cores, remember that some of the steps performed by crossMining consume a great amount of resources due to their complexity – especially the creation of statistical lexica (see page 17). Therefore, selecting all cores is only recommended if the computer does not execute any other (important) programs or processes while creating the lexicon.

The server you are currently connected to is displayed at the bottom of the dialog.

4.2 Advanced Settings

**Number of Iterations**

In the advanced options, you can first determine the number of iterations in the first and second phases of the lexicon creation. (It is difficult to make a general recommendation concerning the optimum number of iterations, as this differs from case to case. The value depends especially on the selected language pair, but also on the selected language direction. For example, the optimum values for the language direction German-English are usually different from those for the language direction English-German due to the different morphology of the two languages.



Therefore, the preset values in the advanced crossMining settings must be considered as a starting point for test purposes. As you conduct your tests, you should optimize these values for your specific data.

For the second phase of the lexicon creation, you can also determine that it is to be skipped. For test purposes, this may be useful if the results of the first phase are to be checked.

**Iteration**

In numerical mathematics, the repeated application of the same computing procedure is referred to as "iteration". The results of an iteration step are used as starting values for the next step in order to get closer and closer to a satisfactory final result.

**Phase 1 vs. Phase 2**

In the first phase of the lexicon creation, the probability of word equivalents is analyzed. In the subsequent second phase, the results of the first phase are further processed. In this context, especially the position of the words is included in the probability calculation.

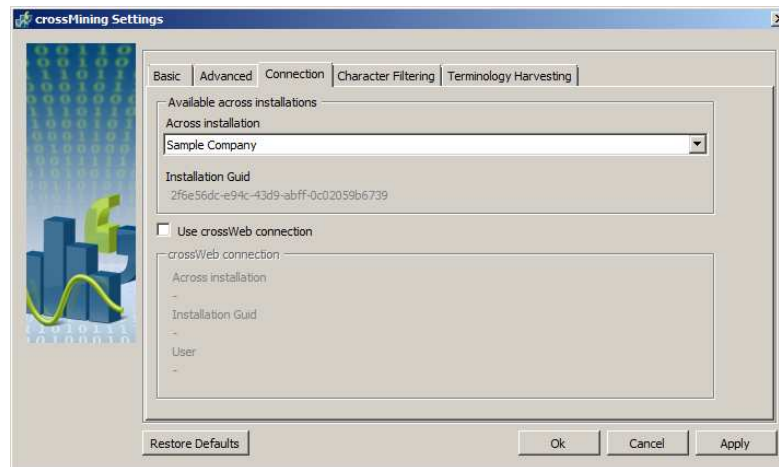
4 Settings

Connection

Moreover, you can specify the minimum probability and the minimum frequency. In this way, you can determine the probability and frequency threshold from which the equivalents are to be written to the statistical lexica.

If the option "Save intermediate results" is enabled, the results of the analyses will be saved after every iteration. This option is only relevant for tests or optimizations by the Across Professional Services team or for inquiries to be sent to the Across Support.

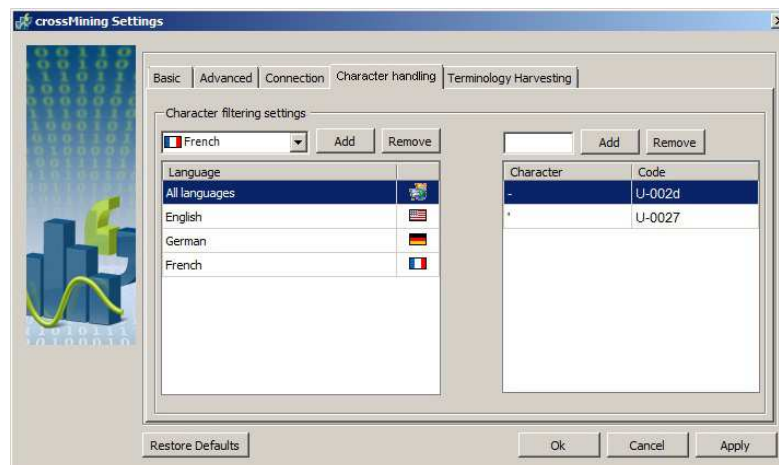
4.3 Connection



In the connection settings, you can determine which Across Server crossMining is to connect to upon start-up. Possible servers are listed in a drop-down list. (To add other servers, generate a generic softkey – see page 9 – and register it on your computer with a double click.)

If you work directly on the computer on which the Across Server is installed, you can also use the connection settings stored for crossWeb. In this case, the user settings entered for the crossAPI Interactive user in crossAdmin under >>**crossWeb** will be used. In this way, it is not necessary to log in separately after starting crossMining. To use the crossWeb connection, activate the corresponding checkbox.

4.4 Character Handling



By default, crossMining filters all special characters such as apostrophes, hyphens, and punctuation marks from the determined source and target-language equivalents. The character handling settings enable the definition of special characters not to be removed by crossMining.



Special characters can be defined globally for all languages or specifically for individual languages. To add characters for individual languages, select the desired language and click **Add**.

Adding characters

To add a special character not to be removed by crossMining, select a language or "All languages" in the left part of the dialog window. Insert the character in the input field in the right part of the dialog window and click **Add**.



To add special characters, you can insert the actual characters or the corresponding Unicode value – introduced by the character string **U-** – e.g. **U-0027** for an apostrophe.



The characters contained in the "All languages" section apply to all languages and may be complemented with the language-specific characters.

Example

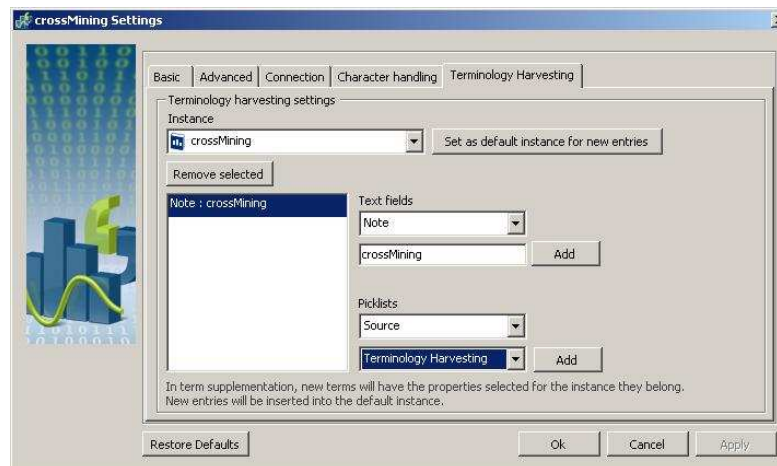
An English-German lexicon is to be created. The following characters are defined under "Character handling":

- All languages: ' -
- English: !
- German: ;

In the English texts of the lexicon, the following characters are retained during the creation of the lexicon: ' - !

In the German texts of the lexicon, the following characters are retained: ' - ;

4.5 Terminology Harvesting



In the terminology harvesting settings, you can determine what is to be done with new entries or terms created in crossTerm within the scope of the terminology harvesting (see page 24).

You can first define an instance as default instance for the bilingual term extraction (see page 27). The new entries created within the scope of the bilingual term extraction will be created in this default instance.

For the creation of terms, for bilingual term extraction , and for the addition of target-language terms (see page 24), you can also determine the text fields and picklist values to be used for the terms.

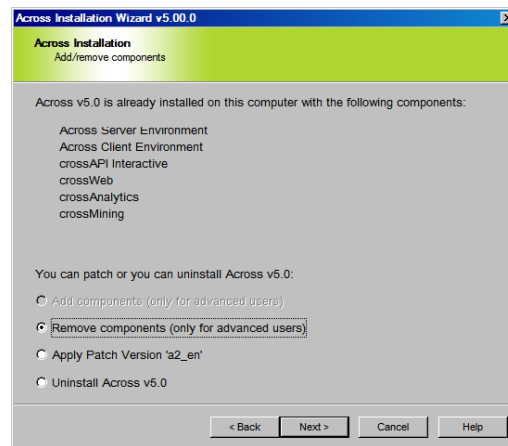
To do this, select a text field or picklist from the list of existing text fields/picklists. In the case of a text field, you can enter the desired text in the input field. In the case of a picklist, you can select the desired value from the list. Click **Add** to add the text field or picklist value.

5 Uninstalling

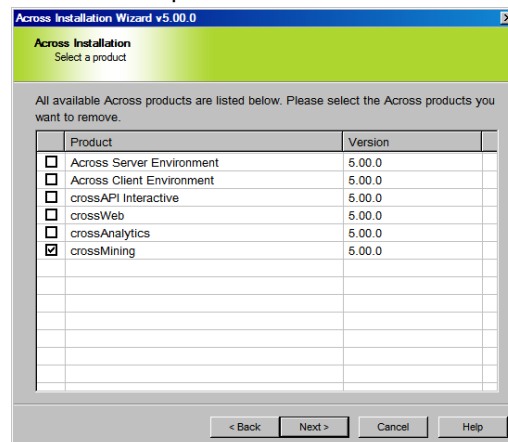
Proceed as follows to uninstall crossMining:

1. The best way to uninstall is by running the **setup.exe**, which you have already run for installing crossMining.
2. Once the wizard has started, click **Next>**.
3. Confirm that you have read the information and wish to uninstall Across or crossMining. Then click **Next >**.
4. Mark the checkbox to confirm that you have read the license agreement (EULA) and accept it. Then click **Next >**.
5. Installed Across components are automatically detected and displayed.

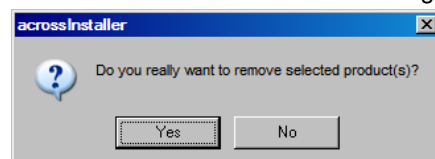
To uninstall all Across components, select the option "Uninstall Across v5.0". (➔) In this case, continue with step 7!) If you merely want to uninstall crossMining, select the option "Add/remove components". (This option is recommended for experienced users only!) Then click **Next >**.

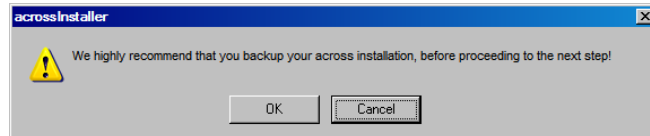


6. Select crossMining by marking the corresponding checkbox. If you are sure that you will also not need crossAPI Interactive, please enable the corresponding checkbox. Make sure that the option "Remove" is enabled. Then click **Next >**.

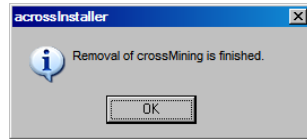


7. Click **Yes** to confirm the two following notifications.





8. crossMining and, if applicable, crossAPI Interactive will now be removed from your computer.
9. Upon completion of the uninstall process, click **OK**.



Uninstalling crossMining via the Control Panel

Instead of uninstalling crossMining via **setup.exe**, you can also uninstall the software in the Control Panel via **>>Start >>Control Panel >>Add/Remove Programs**. Select the entry "crossMining", click **Remove**, and confirm the following message with **Yes**. If you are sure that you no longer need crossAPI Interactive either, proceed in the same way with the entry "crossAPI Interactive".

6 Index

- addition of target-language terms 24
- advanced settings 35
- autocompletion 21
 - in Across 21
 - in crossDesk 21
 - test 22
- bilingual term extraction 27
- creating generic softkey 9
- crossAPI Interactive 6
 - creating generic softkey 9
- crossMining
 - close 33
 - icon toolbar 17
 - install 6
 - introduction 3
 - settings 34
 - advanced 35
 - basic 34
 - character handling 36
 - connection 36
 - terminology harvesting 37
 - start 15
 - uninstall 39
 - use 17
- default output directory 21
- graphs 20
- import of the Moses SMT phrase tables 30
- installation
 - crossAPI Interactive 6
 - crossMining 12
- Moses SMT 30
- process graphs 20
- settings 34
- statistical lexica 17
 - create 18
 - default output directory 21
 - deployment 21
 - crossGrid 21
 - crossWAN classic 21
 - crossWAN load 21
 - online clients 21
 - from Moses SMT phrase tables 30
- statistical lexicons
 - graphs 20
- system requirements 6
- terminology harvesting 24
 - addition of target-language terms 24
 - bilingual term extraction 27