# Khresmoi – Czech IE module – Installation and Usage

## Table of contents:



# 1 Introduction

This document provides installation and usage notes to the Czech information extraction module of the Khresmoi project.

Additional information about the module can be found at the project wiki:
http://wiki.khresmoi.eu/index.php5/Czech_IE_module

# 2 Requirements

Linux is the recommended platform for TectoMT tools. All other components are written in Java and they are platform independent.

## 2.1 Required software components:

- **GATE**
  http://gate.ac.uk/
- **TectoMT**
  http://ufal.mff.cuni.cz/tectomt/
  - including **TrEd**
    http://ufal.mff.cuni.cz/~pajas/tred/
  - and following **blocks** have to be installed:
    - SCzechW_to_SCzechM::TextSeg_tokenizer_and_segmenter
    - SCzechW_to_SCzechM::TagMorce
- **Czsem Mining Suite**
  http://czsem.berlios.de

# 3 Installation

The installation consists of three steps:

## 3.1 Installation of Required Software Components

First of all, it is necessary to install all required software components mentioned above.

Installation of GATE is strait forward and well documented, see the GATE homepage for details.

Installation of TectoMT can be complicated, follow instructions given in Czsem installation instructions and on the TectoMT homepage. After TectoMT installation is complete, ensure that **the two required blocks** (above) are installed too. If not, install them manually according to instructions given in corresponding directories (see in "tectomt\libs\packaged\").

Czsem Mining Suite installation instructions can be found at:
http://czsem.berlios.de/czsem_install.html
Install Czsem using **czsem-1.1.0-standard.jar** (platform independent installer).
Following options have to be selected and properly configured during the Czsem installation:

- Gate integration
- TectoMT integration

## 3.2 Obtain Necessary Resources

The module uses several external resources; part of them is restricted to the usage within the project only and therefore they cannot be publicly distributed.

There are two kinds of the resources:

### 3.2.1 Primary Resources

Primary resources are resources that were obtained directly from corresponding providers. Usage of these resources is restricted to the project only.

They are:

- Czech translation of MeSH (in XML format) and

- The BMČ database of article references (in UNIMARC ISO 2709 format)

These resources are not necessary for the function of the module because the majority of functions do not rely on these resources and the secondary resources should be sufficient. We do not provide download links for these resources; contact the CUNI partner if you need them.

### 3.2.2 Secondary Resources

Secondary resources were created and developed as a part of the module. Some of them represent a condensed version of the primary resources and therefore they cannot be publicly distributed. They can be downloaded from the project wiki:
http://wiki.khresmoi.eu/images/a/af/Czech_IE_resources.zip

Download these resources and unpack them directly to the folder where Czsem Mining Suite was installed (the default name of the folder is "czsem_suite_1.1.0").

## *3.3 Download and Setup Binaries*

Download the Czech IE module binaries from:
http://prdownload.berlios.de/czsem/Khresmoi.Czech_IE-1.0-zip-with-dependencies.zip

Unpack them directly to the Czsem folder (the same as for resources).

Change working directory to the Czsem folder and run KhresmoiConfig using following command:

```
java -jar Khresmoi.Czech_IE-1.0.jar cuni.khresmoi.KhresmoiConfig
```

It should produce a new file "khresmoi_config.xml" in the Czsem folder. This file contains the whole configuration of the module. The configuration has to be adjusted. Open the file and edit corresponding values (file and directory paths) according to your needs. It is necessary to modify the absolute paths; the relative ones should be ok in the default setting (although they can be modified as well). Description of individual values can be found in the following section.

# 4 Usage

The main JAR archive of the module (Khresmoi.Czech_IE-1.0.jar) is configured such that it is possible to use the "java -jar" command to run an arbitrary class form the archive. This is helpful because Java class path is configured inside the JAR and it is not necessary to do any additional class path configuration on the hosting computer.

We strongly recommend to increase the default maximum memory heap size of Java virtual machine, e.g. to run all functions of the module with the "-Xmx1500m" Java switch. It allows Java virtual machine to allocate up to 1.5 GB of memory, which should be sufficient for annotation of larger files. A universal command to run any function of the module looks like:

```
java -Xmx1500m -jar Khresmoi.Czech_IE-1.0.jar <class name>
```

<class name> stands for the name of the class representing given function.

Some of the functions are equipped with so called StopRequestDetector. Such functions can be safely terminated at any time by typing "stop<Enter>" on the terminal. For these functions it is also necessary to type this termination command at the end of processing. We are sorry for this little inconvenience.

List of all classes equipped with StopRequestDetector:

cuni.khresmoi.MorceBatchAnalysis
cuni.khresmoi.InformationExtractionAnalysis
cuni.khresmoi.CompoundAnalysis
cuni.khresmoi.mimir.MimirIndexFeeder
cuni.khresmoi.bmc.BMCStatistics
cuni.khresmoi.PlainTextExport

List of all important classes and their functions follows:

## 4.1 cuni.khresmoi.KhresmoiConfig

It produces "khresmoi_config.xml" file with default configuration in the working directory. Warning: if the "khresmoi_config.xml" file already exists, it will be replaced.

## 4.2 cuni.khresmoi.bmc.BMCGateCorpusBuider

It exports all URLs form the BMČ database and downloads corresponding articles as GATE documents.
**Input**: KhresmoiConfig.bmcIsoFilePath
**Output**: KhresmoiConfig.outputDirBmcOrig

## 4.3 cuni.khresmoi.bmc.BMCGateCorpusFilter

It does filtering of the corpus.
**Input**: KhresmoiConfig.outputDirBmcOrig,
KhresmoiConfig.SerializedResourcesDir \bmcDB.ser
**Output**: KhresmoiConfig.outputDirBmcFilterInlude,
KhresmoiConfig.outputDirBmcFilterExclude

## 4.4 cuni.khresmoi.MorceBatchAnalysis

It runs the linguistic analysis. This class was developed to avoid memory leaks in TectoMT that can be met at certain occasions.
**Input**: KhresmoiConfig.outputDirBmcFilterInlude
**Output**: KhresmoiConfig.outputDirBmcAnalyzedTmp_Morce

## 4.5 cuni.khresmoi.InformationExtractionAnalysis

It runs information extraction analysis. The actually executed GATE application can be changed using the KhresmoiConfig.gateAppIeAnalysis option.
**Input**: KhresmoiConfig.outputDirBmcAnalyzedTmp_Morce,
KhresmoiConfig.gateAppIeAnalysis
**Output**: KhresmoiConfig.outputDirBmcAnalyzed

## 4.6 cuni.khresmoi.CompoundAnalysis

It is very similar to cuni.khresmoi.InformationExtractionAnalysis, but additionally (for space saving reasons) it removes all existing annotations from the analyzed documents. This is not suitable, if these annotations are further needed (e.g. for Mimir indexing). Compound analysis can be executed by cuni.khresmoi.InformationExtractionAnalysis as well, without this side effect.
**Input**: KhresmoiConfig.outputDirBmcAnalyzed,
KhresmoiConfig.gateAppCompoundAnalysis
**Output**: KhresmoiConfig.outputDirBmcAnalyzedCompound

### 4.7 cuni.khresmoi.mimir.MimirIndexFeeder

It sends all GATE documents in given directory to given Mimir index URL.
**Input**: KhresmoiConfig.inputDirMimirIndexFeeder
**Output**: KhresmoiConfig.mimirIndexUrl

### 4.8 cuni.khresmoi.bmc.LinkExport

It exports all article source URLs from GATE documents saved in given directory and then it exports all article source URLs from the whole BMČ database.
**Input**: KhresmoiConfig.inputDirLinkExport, KhresmoiConfig.bmcIsoFilePath
**Output**: bmc_links_filered.txt, bmc_links_all.txt

### 4.9 cuni.khresmoi.PlainTextExport

It produces plain text export to a very simple format: <token>|<lemma>|<m-tag>. Each line represents one sentence.
**Input**: KhresmoiConfig.outputDirBmcAnalyzed
**Output**: KhresmoiConfig.outputDirBmcPlainTextExport

### 4.10 cuni.khresmoi.mesh.MeshStatistics

It produces statistics about English and Czech MeSH terminology.
**Input**: KhresmoiConfig.meshXmlFilePath
**Output**: stderr

### 4.11 cuni.khresmoi.bmc.BMCStatistics

It produces evaluation statistics against BMČ gold standard database for all GATE documents in given directory.
**Input**: KhresmoiConfig.inputDirStatistics,
KhresmoiConfig.SerializedResourcesDir \bmcDB.ser
**Output**: stderr

### 4.12 cuni.khresmoi.bmc.FileInfo

It prints basic BMČ information about individual documents in given directory.
**Input**: KhresmoiConfig.inputDirFileInfo, KhresmoiConfig.bmcIsoFilePath
**Output**: stderr

### 4.13 cuni.khresmoi.mesh.MeshAnnieGazeteer

It produces the Czech MeSH gazetteer list with lemmatization.
**Input**: KhresmoiConfig.meshCzLammatizationAnalysisOutput,
KhresmoiConfig.serializedResourcesDir /meshDB.ser
**Output**: KhresmoiConfig.gazetteerResourcesDir /meshcz_lemmas.lst

### 4.14 cuni.khresmoi.mesh.CompoundGazeteer

It produces two Czech MeSH gazetteer lists using compound analysis and lemmatization.
**Input**: KhresmoiConfig.meshCzLammatizationAnalysisOutput,
KhresmoiConfig.serializedResourcesDir /meshDB.ser
**Output**: KhresmoiConfig.gazetteerResourcesDir /meshcz_lemmas_compound_short.lst,
KhresmoiConfig.gazetteerResourcesDir /meshcz_lemmas_compound_long.lst

### 4.15 Additional Classes for Secondary Resources Construction

There are two additional classes producing secondary resources for the primary ones. They are needed if your Java virtual machine uses a different serialization (java.io.ObjectOutputStream) format than that one used by us for building the provided secondary resources. Following table describes these classes:

| Class name | Input resource | Output resource |
|---|---|---|
| cuni.khresmoi.mesh.MeshRecordDB | meshXmlFilePath | meshDB.ser |
| cuni.khresmoi.bmc.BMCDatabase | bmcIsoFilePath | bmcDB.ser |

# 5 Source Codes Compilation

In cases you need to compile the module from sources we recommend to use Apache Maven: http://maven.apache.org/

Alternatively any Java 1.6 compiler should be sufficient.

Packed sources can obtained from following Maven repository: http://czsem.berlios.de/maven2/repo/

The module artifact id is: “Khresmoi.Czech_IE” and group id is: “cuni”. Current version is “1.0”.

The “Khresmoi.Czech_IE-1.0.pom” file contains Maven configuration: http://czsem.berlios.de/maven2/repo/cuni/Khresmoi.Czech_IE/1.0/Khresmoi.Czech_IE-1.0.pom

All necessary libraries (jars) are packed in the main binaries archive (already mentioned above): http://prdownload.berlios.de/czsem/Khresmoi.Czech_IE-1.0-zip-with-dependencies.zip

Some of the libraries have to be installed manually to the local Maven repository (mvn install:install-file).

The sources are available also at the public subversion repository:

```
svn checkout svn://svn.berlios.de/czsem/trunk/src/java/khresmoi/
```