KCONNECT

www.kconnect.eu

Toolkit and Report for Translator Adaptation to New Languages (First Version)

Deliverable number	*D1.2*
Dissemination level	*Public*
Delivery date	*1 November 2015*
Status	*Final*
Author(s)	*Aleš Tamchyna, Jindřich Libovický, Pavel Pecina*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644753 (KConnect)

Executive Summary

This report presents the first version of the toolkit for adaptation of the Khresmoi Translator to support translation between new language pairs. The machine translation (MT) system developed within the Khresmoi project supports translation of medical search queries in Czech, German, and French into English and translation of full sentences (from the medical domain) from English into the same languages (Czech, German, and French). The Khresmoi Translator is a statistical phrase-based system built around Moses, the state-of-the-art statistical machine translation decoder, and various other tools, including MT Monkey which was developed within Khresmoi to allow the system to be employed as a webservice.

The first version of the toolkit for adaptation of the Khresmoi translator to new languages focuses mostly on simplification of the training procedure which produces statistical models and parameter settings which then can be plugged into the Khresmoi Translator. This is done through Eman Lite, a newly developed set of scripts for training the models and tuning the system parameters. In this report we present a user documentation of Eman Lite and also a collection of data that can be used to train a medical-domain translation system for English-Czech, English-German, English-French and newly also English-Hungarian, English-Polish, English-Spanish, and English-Swedish.

Table of Contents

List of Abbreviations

(S)MT	(Statistical) Machine Translation
TM	Translation Model
LM	Language Model
EN	English
CS	Czech
DE	German
HU	Hungarian,
FR	French
PL	Polish
ES	Spanish
SV	Swedish

1 Introduction

Khresmoi Translator, the machine translation (MT) system developed within the Khresmoi project allows translation of medical queries from non-English languages into English and translation of full sentences from English into the same non-English languages. The non-English languages which are already supported include Czech, German, and French. In the KConnect project, we will adapt the system to support for four new languages, including Hungarian, Polish, Spanish and Swedish. This adaptation will be realized using a newly developed toolkit, which will simplify the entire process as much as possible.

The Khresmoi Translator is based on Moses [1], a state-of-the-art toolkit for Statistical Machine Translation (SMT), and a set of other tools for preprocessing the input text and postprocessing its translation. The system is deployed as a webservice, which is enabled through MTMonkey, a software designed and developed in Khresmoi.

The Khresmoi Translator is statistical. It has several components (models) which need to be trained on a collection of training data (parallel and monolingual). The first version of the adaptation toolkit focuses mostly on simplification of the training procedure which produces the statistical models and parameter settings to be plugged into the Kreshmoi Translator. This procedure is implemented in Eman Lite, a newly developed set of scripts for training the models and tuning the system parameters.

In Section 2 of this report, we present the user documentation of Eman Lite and in Section 3, we review a collection of data that can be used to train a medical-domain translation system for English-Czech, English-German, English-French and newly also English-Hungarian, English-Polish, English-Spanish, and English-Swedish.

2 Eman Lite - a Toolkit for Automated Training of SMT Systems

2.1 Installation

The recommended platform for this toolkit is Ubuntu 14.04 LTS. However, Eman Lite should work on any sensible Linux distribution.

The following is a list of tools required to run Eman Lite. The corresponding Ubuntu package is listed for each item.

• GCC, autotools, make, libtool and other build tools `[build-essential]`

• CMPH + header files `[libcmph-dev]`

• LibXMLRPC + header files `[libxmlrpc-core-c3-dev, libxmlrpc-c++8-dev]`

• git `[git]`

• bzip2 + header files `[libbz2-dev]`

• Boost libraries `[libboost-all-dev]`

• Python Regex module `[python-regex]`

Once the prerequisites are met, Eman Lite can be installed by running the following command in the Eman Lite directory:

```
./prepare-environment.sh [installation-directory]
```

By default, the installation script will build everything in the directory install. The script downloads and compiles all tools required for training a Moses-based SMT system, namely:

Moses

An SMT decoder and toolkit (includes KenLM).

GIZA++

An implementation of IBM models for word alignment.

SALM

An efficient implementation of suffix arrays.

Upon successful completion, the installation script writes the installation directory in the file `prepare-environment.OK`.

The installation can further be tested by running the script `test-environment.sh` (without any arguments). This test builds a small MT system from the data sample distributed along with Eman Lite and checks whether it performs as expected.

2.2 Building an SMT System

Eman Lite is run using the command `train-system.sh`. Running the script with the option `--help` outputs the following message:

```
Usage: train-system.sh [options] working-directory parallel-source parallel-target

  Arguments:

    working-directory : where train-system.sh will work, cannot be non-empty

    parallel-source : source side of parallel corpus

                      (plain text, one sentence per line)

    parallel-target : target side of parallel corpus

                      (same number of lines as source)

  Accepted options:

    -h|--help : print this help message

    --mono=MONO : file with additional monolingual data

    --dev-src=DEV-SRC --dev-tgt=DEV-TGT : use jointly to provide a development set

    --test-src=TEST-SRC --test-tgt=TEST-TGT : use jointly to provide a test set
```

To train a translation system, the user must provide at least the mandatory (positional) arguments: the target working directory, the source side of the parallel training data and the target side of the parallel data.

All data are assumed to be plain text files encoded in UTF-8. Other formats (XML, gzip,...) and other encodings are not supported. It is up to the user to provide all data files in the required form.

Both mandatory and optional data files are described in this section.

Parallel training data

The parallel corpus is the essential ingredient for training a statistical machine translation (SMT) system. The user must provide two files. The first file should contain sentences in the source language, one sentence per line. The second file should contain their translations on the corresponding lines. Both files should therefore have the same length. See the files `test/train.src` and `test/train.tgt` for inspiration.

The target side of the parallel data is automatically used also for language model training. The user is free to add more monolingual data using the option `--mono`.

If not provided by the user, the development and test data are extracted from these files.

[Optional] Monolingual training data

Additional monolingual data (in the target language) can be provided to improve the fluency of produced translations. The data should be contained in a single plain-text file with sentences separated by line breaks. The file path is passed to train-system.sh using the option `--mono`.

[Optional] Development set

The development set is used to optimize the parameters of the MT system. The user does not have to provide the development data, in which case Eman Lite reserves a small portion of the parallel training data for this purpose. The parameters `--dev-src` and `--dev-tgt` must be used jointly to specify the files containing the source and target side of the development data. Sentences in the files must correspond to each other line by line.

Common size of the development data is several thousand sentence pairs. Ideally, the data should be as similar as possible (in terms of genre, vocabulary) to the texts that will be translated by the final system.

[Optional] Test set

As a final step, the full MT system is evaluated on an independent, unseen data set to estimate its expected performance. Similarly to the development data, this set is optional. The files are specified using the options `--test-src` and `--test-tgt`. Sentences in the files must correspond to each other line by line.

2.3 Configuration

Eman Lite is configured by modifying the global configuration file `CONFIG.sh`. This section explains the individual settings.

CORES

How many CPU cores to use.

DEVSIZE

Size of development data in sentences. This option only applies when the user does not provide their own development corpus. In this case, development data of length `DEVSIZE` are extracted from the provided parallel corpus.

TESTSIZE

Size of test data in sentences. Similarly to the development data, this option only applies when a separate test set is not specified.

MINPARASIZE

Minimal accepted size of parallel data (in sentence pairs). This serves as a simple check that training the MT system is sensible.

MAXBLOCKSIZE

How many parallel sentences can be processed in one block. When the parallel data size exceeds this value, the training data is split into multiple parts. A separate translation model is extracted from each part and the models are later combined. This approach ensures that training time remains reasonable when the provided data is large and that the machine does not run out of memory.

LMORDER

The order of the language model, i.e. how long sequences of words are captured.

LMARGS

Arguments for building the language model. The default setting should be adequate for most cases. Pruning (by absolute discounting) can be applied by changing the settings here. Different values can have significant impact of the trade-off between model size/speed and translation quality. See the documentation for `lmplz`.[1]

LMBINARIZEARGS

Arguments for language model binarization. Default settings should be adequate. More details can be found in the documentation of the tool `build_binary`.[2]

1 https://kheafield.com/code/kenlm/estimation/

2 https://kheafield.com/code/kenlm/structures/

SIGFILTERARGS

Arguments for the statistical phrase-table filter. Similarly to pruning settings in `LMARGS`, different values can be utilized to balance the trade-off between speed/size and translation quality. Details of the possible settings can be found in `sigtest-filter` documentation. [1]

2.4 Implementation Details

This section covers some of the more technical topics regarding the implementation of Eman Lite.

2.4.1 Corpus Filter

Eman Lite uses the script filter-parallel.pl to discard sentence pairs which could cause errors in the MT training pipeline. The script is configurable and Eman Lite uses the default setting which includes the following filters:

utf8

Throw away sentences with invalid UTF-8 characters.

escapemoses

Escape characters which are interpreted by Moses (currently only the pipe symbol "|").

normspace

Normalize whitespace to a single space character between tokens.

minwords:1

Throw away sentences with zero tokens.

salmlimits

Throw away sentences which would cause trouble in SALM (containing too many tokens or tokens which are too long).

2.4.2 Segmentation of Parallel Data

In order to speed up training and to reduce the maximum memory consumption, Eman Lite splits parallel data into blocks which are processed independently. The default size of these blocks is 1 million sentence pairs. Word alignment, phrase extraction and phrase table creation are done separately for each of the blocks. Notably, statistical filtering is also run independently on each of the phrase tables. Finally, linear phrase table interpolation is applied with uniform weights to produce the final translation model. The tool `tmcombine` (bundled with Moses) is used for the interpolation.

1 https://github.com/moses-smt/mosesdecoder/blob/master/contrib/sigtest-filter/

3 Overview of Available Training Data

For current SMT systems, the amount and quality of the training data is the crucial factor for the resulting translation system quality. In general, two types of data required to train an SMT system: parallel data which consist of pairs of sentences in source and target language, and monolingual data which consist of sentences in the target language only. Parallel data are used for training the translation model which produces candidates for phrases translation.

We use a combination of in-domain data and general data. Using the training data from medical texts allows to adapt the translation system to such sentences and to acquire domain-specific vocabulary. Because of the relative sparsity of the in-domain data we mix them with the general domain data and thus increase robustness of both the translation and language model.

The most prominent data source for MT is EU legislation and other documents which are published in all official languages of EU. Other important resources are film subtitles and web crawls which can be easily obtained automatically from the Internet. An important source for German and French are also patent applications. For a detailed overview of the data sources, see the next section which contains also summaries of legal status of each dataset which limits its usage.

The overview of the training data size is provided in Table 1. Note that the numbers are only indicative since some of the data sources overlap and thus do not bring so much new information to the models.

	parallel data with English		monolingual data	
	in-domain	general	in-domain	General
Spanish	76	1420	2	474
Czech	19	664	20	698
German	128	309	355	689
French	220	931	927	1451
Hungarian	18	652	13	98
Polish	17	595	13	205
Swedish	23	402	21	158
English	–	–	4280	3370

Table 1: Overview of training data size (in millions of tokens)

The amount of available training data vary for the languages. In general, the languages with more speakers usual have more data available. This also influences the quality of the translation substantially.

3.1 Parallel Data

This section summarizes the parallel data used from training the translation model of the SMT system. It is a compilation of previously existing data sets, data from the Khresmoi project and data newly collected for the KConnect project. The data size computed in millions of tokens of the English side of the corpora to allow direct comparison of the dataset sizes.

3.1.1 Medical Domain

	ES	CS	DE	FR	HU	PL	SV
EMEA	14.8	14.4	14.9	14.8	14.3	14.3	14.8
PatTR	–	–	102	127	–	–	–
COPPA - in domain	–	–	–	67.2	–	–	–
MuchMore	–	–	1.0	–	–	–	–
UMLS	56.6	6.3	9.4	9.7	3.8	2.7	1.4
Health Web Crawl	3.7	–	–	–	–	–	–
Subtitles	1.7	–	0.8	1.5	–	.03	1.0
MeSH	.12	.10	.11	.15	1.7	–	.56
SnomedCT	–	–	–	–	–	–	5.9
HU Abstracts	–	–	–	–	0.5	–	–
CESTA	–	–	–	0.09	–	–	–
ECDC	0.04	0.04	0.04	0.04	0.04	0.04	0.04
WIKI titles - medical	.01	.004	.01	.01	.002	.01	.005

Table 2: Overview of the sources of parallel training data size from the medical domain (in millions of tokens)

EMEA corpus

A parallel corpus made out of PDF documents from the European Medicines Agency collected by Uppsala University.

Licence: Can be freely used for both commercial and non-commercial purposes, the European Medicines Agency must be always acknowledged.

Languages: ES-EN CS-EN DE-EN FR-EN HU-EN PL-EN SV-EN

URL: http://opus.lingfil.uu.se/

Contact: Jörg Tiedemann (jorg.tiedemann@lingfil.uu.se), Uppsala University

PatTR: Patent Translation Resource

PatTR is a sentence-parallel corpus extracted from the MAREC patent collection. The current version contains more than 22 million German-English and 18 million French-English parallel sentences

collected from all patent text sections as well as 5 million German-French sentence pairs from patent titles, abstracts and claims.

Licence: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Languages: DE-EN, FR-EN

URL: http://www.cl.uni-heidelberg.de/statnlpgroup/pattr/

https://heidata.uni-heidelberg.de/dvn/dv/statnlpgroup/faces/study/StudyPage.xhtml?globalId=doi:10.11588/data/10002

Contact: Katharina Wäschle (waeschle@cl.uni-heidelberg.de), Universität Heidelberg

COPPA

COPPA is the Corpus Of Parallel Patent Applications provided by WIPO (World Intellectual Property Organization) of English French Patent Cooperation Treaty applications (title and abstract) published between 1990 and 2010.

Licence: Free for research purposes, distributed on DVD only.

Languages: FR-EN

URL: http://www.wipo.int/patentscope/en/data/#coppa

Contact: Use contact form on the WIPO website (http://www.wipo.int/contact/en/area.jsp?area=patentscope_fb)

MuchMore Springer Bilingual Corpus

The corpus used in the MuchMore project is a parallel corpus of English-German scientific medical abstracts obtained from the Springer Link web site. The corpus consists approximately of 1 million tokens for each language. Abstracts are from 41 medical journals, each of which constitutes a relatively homogeneous medical sub-domain.

Licence: Licence is unclear. Produced within an EU project MUCHMORE in 2001, the corresponding deliverable (http://muchmore.dfki.de/pubs/D4.1.pdf) does not mention the license issues at all.

Languages: DE-EN

URL: http://muchmore.dfki.de/resources1.htm

Contact: Hans Uszkoreit (Hans.Uszkoreit@dfki.de), DFKI

UMLS (Unified Medical Language System)

A multilingual metathesaurus (including Czech, English, French, and German) of health and biomedical vocabularies and standards.

Licence: Software with UMLS as an integral part can be distributed without limitations. Anyone with the license cannot distribute the data further. The detailed license conditions are here: https://uts.nlm.nih.gov/license.html

Languages: ES-EN, CS-EN, DE-EN, FR-EN, HU-EN, PL-EN, SV-EN

URL: http://www.nlm.nih.gov/research/umls/

Contact: See http://www.nlm.nih.gov/research/umls/support.html

Medical Web Crawl

Parallel texts crawled from two U.S. public web pages (http://www.cancer.gov/, www.nlm.nih.gov/medlineplus) and the World Health Organization. Other languages can be added. Created for the KConnect project.

Licence: Data are crawled directly from webs of public institutions, texts are public domain.

Languages: ES-EN

URL: N/A

Contact: Jindřich Libovický (libovicky@ufal.mff.cuni.cz), Charles University in Prague

Subtitles

Subtitles of recent medical TV series (House M.D., Scrubs, Grye's Anatomy) downloaded from major subtitles providers (Addic7ed, OpenSubtitles, Podnapisi, TheSubDB, TvSubtitles). Czech and Hungarian is not included because of the encoding issues. Created for the Kconnect project.

Licence: Very unclear. The subtitles are freely downloadable, however without permission of the copyright holders.

Languages: ES-EN, DE-EN, FR-EN, PL-EN, SV-EN

URL: N/A

Contact: Jindřich Libovický (libovicky@ufal.mff.cuni.cz), Charles University in Prague

MeSH

Medical Subject Headings (MeSH) is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences; it serves as a thesaurus that facilitates searching. Created and updated by the United States National Library of Medicine (NLM), it is used by the MEDLINE/PubMed article database and by NLM's catalog of book holdings. MeSH is also used by ClinicalTrials.gov registry to classify which diseases are studied by trials registered in ClinicalTrials.gov.

Licence: Khresmoi languages extracted from UMLS (https://uts.nlm.nih.gov/help/license/LicenseAgreement.pdf)

Sv license owned by Findwise, Hu license owned by Precognox.

Languages: ES-EN, CS-EN, DE-EN, FR-EN, HU-EN, SV-EN

URL: http://www.ncbi.nlm.nih.gov/mesh

sv: http://mesh.kib.ki.se/swemesh/licence_en.html

Contact: Use form on the MeSH website.

Swedish SnomedCT

SNOMED Clinical Terms is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. Provided to the KConnect project by Findwise.

Licence: Licence owned by Findwise

Languages: SV-EN

URL: N/A

Contact: Fredrik Axelsson (fredrik.axelsson@findwise.com), Findwise

Hungarian Abstracts

Abstracts of Hungarian medical papers from Napivizit.hu which is owned by our partner, Akademiai Kiado (Academic Press). Collected by Precognox for the KConnect project.

Licence: License owned by Akademiai Kiado (Academic Press)

Languages: HU-EN

URL: http://www.napivizit.hu/

Contact: Zoltan Varju (zvarju@precognox.com), Precognox

ELRA-E0020, CESTA Evaluation Package

Subpart: English-French parallel corpus from the second campaign data. Includes an adaptation corpus of 19,383 English words and 22,741 French words + a test corpus of 18,880 English words and 23,411 French words.

Licence: Data published by European Language Resource Association. Charles University owns a non-commercial licence based on its ELRA membership.

Languages: FR-EN

URL: http://catalog.elra.info/product_info.php?products_id=994

Contact: Khalid Choukri (choukri@elda.org)

ECDC

In October 2012, the European Union (EU) agency 'European Centre for Disease Prevention and Control' (ECDC) released a translation memory (TM), i.e. a collection of sentences and their professionally produced translations, in twenty-five languages. The data gets distributed via the web pages of the EC's Joint Research Centre (JRC). Here we describe this resource, which bears the name ECDC Translation Memory, short ECDC-TM.

Licence: The data is property of the European Commission. The licence allows both commercial and non-commercial use of the data.

Languages: CS-EN, DE-EN, FR-EN, ES-EN, SV-EN, PL-EN, HU-EN

URL: https://ec.europa.eu/jrc/en/language-technologies/ecdc-translation-memory

Contact: No person mentioned, only email is provided (webmaster@ecdc.europa.eu)

Wikipedia titles

Titles of Wikipedia articles automatically extracted from DBpedia. The in-domain titles are extracted using a handcrafted list of medical categories (provided by Medical University of Vienna) in English and the interlingual links. The dataset was originally created for the Khresmoi project, now it has been extended for the new languages.

Licence: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License (as everything that comes from Wikipedia)

Languages: ES-EN, CS-EN, DE-EN, FR-EN, HU-EN, PL-EN, SV-EN

URL: N/A

Contact: Jindřich Libovický (libovicky@ufal.mff.cuni.cz), Charles University in Prague

3.1.2 General Domain

	ES	CS	DE	FR	HU	PL	SV
WIKI titles - general	1.7	.4	1.9	2.3	0.2	1.5	0.8
UN - general	241	–	5.7	143	–	–	–
EU Bookshop	14.2	11.6	23.5	26.6	12.2	14.0	51.9
Europarl	56.4	17.6	55.0	57.7	17.0	17.2	52.5
JRC-Acquis	18.4	30.0	18.5	18.5	12.6	37.8	18.0
Opensubtitles	912	532	134	574	510	447	206
News Commentary	4.7	3.4	4.8	4.4	–	–	–
Hunglish	–	–	–	–	29	–	–
DGT-Acquis	173	139	134	170	142	142	148
Cordis-Rapid	–	–	–	–	–	7.2	–
Hansard	–	–	–	20.2	–	–	–
COPPA - all	–	–	–	465	–	–	–
Linguee	–	–	.07	–	–	–	–

Table 3: Overview of the sources of parallel training data size from general domain (in millions of tokens)

Wikipedia titles

See in-domain data.

MultiUN: Multilingual UN Parallel Text 2000—2009

The MultiUN parallel corpus is extracted from the United Nations Website, and then cleaned and converted to XML at Language Technology Lab in DFKI GmbH (LT-DFKI), Germany. The documents were published by UN from 2000 to 2009. The texts are from general domain.

Licence: Neither the paper published about the corpus, nor the corpus web page mention any licensing issues. The corpus is therefore under the same licence as the UN materials themselves. The corpus was gathered within an EU-funded project EuromatrixPlus.

Languages: ES-EN, DE-EN, FR-EN

URL: http://www.euromatrixplus.net/multi-un/

Contact: Andreas Eisele (Andreas.Eisele@dfki.de), DFKI

EU Bookshop

The corpus was compiled from the EU Bookshop website – an online service and archive of publications from various European institutions. The service contains a large body of publications in the 24 official languages of the EU.

Licence: Texts are owned by the EU institutions. Free for commercial and non-commercial when properly acknowledged.

Languages: CS-EN, DE-EN, FR-EN, ES-EN, SV-EN, PL-EN, HU-EN

URL: http://opus.lingfil.uu.se/EUbookshop.php

Contact: Jorg Tiedemann (jorg.tiedemann@lingfil.uu.se), Uppsala University

Europarl

The Europarl parallel corpus is extracted from the proceedings (mostly speeches) of the European Parliament.

Licence: The authors are not aware of any copyright restrictions of the material. If you use this data in your research, please contact the authors.

Languages: CS-EN, DE-EN, FR-EN, ES-EN, SV-EN, PL-EN, HU-EN

URL: http://www.statmt.org/europarl/

Contact: Philipp Koehn (pkoehn@inf.ed.ac.uk), University of Edinburgh and Johns Hopkins University

JRC-Acquis

JRC-Acquis is a collection of legislative text of the European Union and currently comprises selected texts written between the 1950s and now.

Licence: Texts are owned by the EU institutions. Free for non-commercial use when properly acknowledged.

Languages: CS-EN, DE-EN, FR-EN, ES-EN, SV-EN, PL-EN, HU-EN

URL: http://opus.lingfil.uu.se/JRC-Acquis.php

Contact: Jorg Tiedemann (jorg.tiedemann@lingfil.uu.se), Uppsala University

Opensubtitles

A collection of documents from http://www.opensubtitles.org/.

Licence: The data ownership is unclear. Opensubtitles require to place a link to their website whenever the data are used.

Languages: CS-EN, DE-EN, FR-EN, ES-EN, SV-EN, PL-EN, HU-EN

URL: http://opus.lingfil.uu.se/OpenSubtitles.php

Contact: Jorg Tiedemann (jorg.tiedemann@lingfil.uu.se), Uppsala University

News Commentary

The News Commentary parallel corpus comprises news and commentary texts from publicly available sources provided by organizers of the series of Workshops on Machine Translation as shared task training data. The WMT 11 version is available for the following language pairs: FR–EN, DE–EN, and CS–EN.

Licence: Licence unknown. Potential users are asked to contact the author.

Languages: CS-EN, DE-EN, FR-EN, ES-EN

URL: http://statmt.org/wmt11

Contact: Philipp Koehn (pkoehn@inf.ed.ac.uk), University of Edinburgh and Johns Hopkins University

Hunglish

The Hunglish Corpus is a free sentence-aligned Hungarian-English parallel corpus of about 120 million words in 4 million sentence pairs. This is the Version 2.0 release of the Corpus, approximately doubling the size of the original 1.0 release from 2005. The raw corpus was gathered from the Web. It consists of several distinct subcorpora: classical literature, modern literature, legal texts, software documentation, movie subtitles.

Licence: Some raw materials used for the Hunglish corpus are under copyright (modern literature, movie subtitles). The Hunglish Corpus is open for use under the Creative Commons Attributions license.

Languages: HU-EN

URL: http://www.hunglish.hu/

Contact: Dániel Varga (daniel@mokk.bme.hu), Budapest University of Technology and Economics

DGT-Acquis

The DGT-Acquis is a family of several multilingual parallel corpora extracted from the Official Journal of the European Union (OJ), consisting of documents from the middle of 2004 to the end of 2011 in up to 23 languages. It partially overlaps with other corpora based from EU data.

Licence: Texts are owned by the EU institutions. Free for both commercial and non-commercial use when properly acknowledged.

Languages: CS-EN, DE-EN, FR-EN, ES-EN, SV-EN, PL-EN, HU-EN

URL: https://ec.europa.eu/jrc/en/language-technologies/dgt-acquis

Contact: M.T. Carrasco Benitez (manuel.carrasco-benitez@ec.europa.eu) Directorate-General for Translation (DGT)

Cordis-Rapid

Automatically aligned bilingual texts crawled from the CORDIS news database and RAPID press releases of the EU.

Licence: Creative Commons Attribution 3.0 Unported Licence

Languages: PL-EN

URL: http://pelcra.pl/new/engpol_18

Contact: PELCRA (Polish and English Language Corpora for Research and Applications) group at the University of Łódź, (contact@pelcra.pl)

Hansard French/English Corpus

The Hansard Corpus consists of parallel texts in English and Canadian French, drawn from official records of the proceedings of the Canadian Parliament. While the content is therefore limited to legislative discourse, it spans a broad assortment of topics and the stylistic range includes spontaneous discussion and written correspondence along with legislative propositions and prepared speeches.

Licence: Charles University owns a licence based on its LDC membership. For external uses LDC User Agreement for Non-Members holds.

Languages: FR-EN

URL: https://catalog.ldc.upenn.edu/LDC95T20

Contact: LDC office (ldc@ldc.upenn.edu)

COPPA

See in-domain data.

Linguee

Linguee is a general human-readable dictionary which provides part of the dictionary for download.

Licence: GPL Version 2 or later; GNU General Public License.

Languages: DE-EN

URL: http://www.linguee.de/

Contact: Unknown

3.2 Monolingual Data

This section summarizes the monolingual data collected for training the language model for the SMT system. The dataset sizes are computed millions of tokens in the language of the dataset, so the numbers are not directly comparable. Note the in addition to this data the target language side of the parallel data can be also used for training the language model.

3.2.1 Medical Domain

	ES	CS	DE	FR	HU	PL	SV	EN
SV Med. Journals	–	–	–	–	–	–	20	–
BMČ	–	16.7	–	–	–	–	–	–
Radio2wiki	–	–	.10	–	–	–	–	–
CESART	–	–	–	11.4	–	–	–	–
EqueR	–	–	–	13.4	–	–	–	–
Precognox crawl	–	–	–	–	12	12	–	–
CLEF 2014	–	–	–	–	–	–	–	685
Trip	–	–	–	–	–	–	–	2120
BioMedCentral	–	–	–	–	–	–	–	367
Cochrane	–	–	–	–	–	–	–	67
Drugbank	–	–	–	–	–	–	–	.89
FMA	–	–	–	–	–	–	–	.88
Genia	–	–	–	–	–	–	–	.56
HON	–	3.2	351	900	–	–	–	1158
Wikipedia med.	1.7	0.6	4.2	2.1	0.6	0.6	0.8	0.5

Table 4: Overview of the sources of monolingual training data size from medical domain (in millions of tokens)

Swedish Medical Journals

Texts from Läkartidningen - a web server for with medical news from years 1996-2009.

Licence: No licence information available.

Languages: SV

URL: http://spraakbanken.gu.se/eng/resources/corpus

Contact: Department of Swedish at the University of Gothenburg (sb-info@svenska.gu.se)

Bibliographia medica Čechoslovaca

A database of all medical professional texts published in Czechoslovakia from 1947.

Licence: Licences signed for internal use only by Charles University.

Languages: CS

URL: http://www.nlk.cz/informace-o-nlk/odborne-cinnosti/bmc

Contact: Národní lékařská knihovna [National Medical Library] (nml@nlk.cz)

Radio2Wiki

Radio2wiki is the text from the german textbook on radiologic diagnosis: Lehrbuch der radiologisch-klinischen Diagnostik by Lechner and Breitenseher.

Licence: Unknown

Languages: DE

URL: http://www.nlk.cz/informace-o-nlk/odborne-cinnosti/bmc

Contact: Franz Kainberger (franz.kainberger@meduniwien.ac.at), Medical University of Vienna

University Publisher 3.0 (office@universitypublisher.com)

CESART Evaluation Package

CESART Evaluation Package was produced within the French national project CESART (Evaluation of terminology extraction tools). Apart from software tools, it contains three domain-specific corpora in French, one of which is this medical corpus.

Licence: Data published by European Language Resource Association. Charles University owns a non-commercial licence based on its ELRA membership.

Languages: FR

URL: http://catalog.elra.info/product_info.php?products_id=993&language=en

Contact: No specific contact provided, ELRA contact form: http://catalog.elra.info/contact_us.php

EQueR Evaluation Package

The corpus contains data from the French Evaluation campaign of question-answering systems. The subpart containing the medical domain is used.

Licence: Data published by European Language Resource Association. Charles University owns a non-commercial licence based on its ELRA membership.

Languages: FR

URL: http://catalog.elra.info/product_info.php?products_id=996

Contact: No specific contact provided, ELRA contact form: http://catalog.elra.info/contact_us.php

Crawled Polish and Hungarian webs

Crawl of Polish and Hungarian medical websites prepared by Precognox.

Licence: The dataset contains only texts that were published on web. For particular license restriction see the websites.

Languages: HU, PL

URL: N/A

Contact: Zoltan Varju (zvarju@precognox.com), Precognox

CLEF

This collection contains documents covering a broad set of medical topics, and does not contain any patient information. The documents in the collection come from several online sources, including the Health On the Net organization certified websites, as well as well-known medical sites and databases (e.g. Genetics Home Reference, ClinicalTrial.gov, Diagnosia). The data were collected for the CLEF eHealth shared task within the Khresmoi project. There is a significant overlap with the HON dataset.

Licence: The dataset was distributed only to the registered participants of the competition. Commercial license can be obtained from ELRA.

Languages: EN

URL: http://catalog.elra.info/product_info.php?products_id=1238

Contact: No specific contact provided, ELRA contact form. http://catalog.elra.info/contact_us.php

Trip

Collection of documents from TRIP, a clinical search engine designed to allow users to quickly and easily find and use high-quality research evidence to support their practice and/or care.

Licence: Full-text database can be obtained after registration. For commercial use a purchased licence from Trip Database is required.

Languages: EN

URL: https://www.tripdatabase.com

Contact: contact@tripdatabase.com, Jon Brassey (jon.brassey@tripdatabase.com)

BioMed Central

A crawl of papers published by BioMed Central. BioMed Central is an STM (Science, Technology and Medicine) publisher of 291 peer-reviewed open access journals. All original research articles published by BioMed Central are made freely accessible online immediately upon publication.

Licence: The Creative Commons CC0 1.0 Public Domain Dedication waiver applies to all published data in BioMed Central open access articles.

Languages: EN

URL:

Contact: Ivan Eggel (ivan.eggel@hevs.ch), Allan Hanbury (allan.hanbury@tuwien.ac.at)

Cochrane

The dataset comprises English reviews of primary research in human healthcare and health policy gathered by the Cochrane - a network of medical researches and professionals.

Licence: For internal use only.

Languages: EN

URL: http://www.cochrane.org/ (dowload: https://download.ir-facility.org/khresmoi/coch.zip)

Contact: Contact form on Cochrane webpage (http://www.cochrane.org/contact)

Drugbank

Bioinformatics and cheminformatics descriptions of drugs in English collected from Drugbank - a richly annotated database of drug and drug target information.

Licence: Can be used commercially after noting the producer and when acknowledged in all materials.

Languages: EN

URL: http://www.drugbank.ca/downloads

Contact: Wishart Research Group, University of Alberta (http://feedback.wishartlab.com/?site=drugbank)

FMA - Foundational Model of Anatomy

The Foundational Model of Anatomy Ontology (FMA) is an evolving computer-based knowledge source for biomedical informatics; it is concerned with the representation of classes or types and relationships necessary for the symbolic representation of the phenotypic structure of the human body in a form that is understandable to humans and is also navigable, parseable and interpretable by machine-based systems.

Licence: Distributed under the Creative Commons Attribution 3.0 Unported License.

Languages: EN

URL: http://sigpubs.biostr.washington.edu/view/projects/Foundational_Model_of_Anatomy.html

Contact: Structural Informatics Group at the Washington University (http://sig.biostr.washington.edu/)

Genia

The GENIA corpus is the primary collection of biomedical literature compiled and annotated within the scope of the GENIA project. The corpus was created to support the development and evaluation of information extraction and text mining systems for the domain of molecular biology

Licence: No licence information available

Languages: EN

URL: http://www.nactem.ac.uk/genia

Contact: Tsujii Laboratory of University of Tokyo (http://www.nactem.ac.uk/tsujii)

HON

HON corpora comprises texts from the medical domain crawled during the Khresmoi project. The source web sites are all HONcode certified and come in a variety of languages. The language of each page is automatically assigned.

Licence: For internal use only, provided by HON.

Languages: CS, DE, EN, FR

Contact: Celia Boyer (Celia.Boyer@healthonnet.org), HON

Wikipedia dumps

Dumps of the wikipedia in the languages we are interested in. The in-domain data were selected using the handcrafted list provided by Medical University of Vienna.

Licence: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License (as everything that comes from Wikipedia)

Languages: CS, DE, EN, ES, FR, HU, PL, SV

URL: N/A

Contact: Jindřich Libovický (libovicky@ufal.mff.cuni.cz), Charles University in Prague

3.2.2 General Domain

	EN	CS	DE	FR	HU	PL	SV	EN
Wikipedia gen.	474	93	699	589	98	205	158	2015
ČNK	–	599	–	–	–	–	–	–
Gigaword	–	–	–	862	–	–	–	1355

Table 5: Overview of the sources of monolingual training data size from general domain (in millions of tokens)

Wikipedia dumps

See in-domain data.

Czech National Corpus

The Czech National Corpus is an academic project focusing on building a large electronic corpus of mainly written Czech. It is primarily designed for linguistic research.

Licence: Charles University has a licence for internal purposes only.

Languages: CS

URL: http://ucnk.ff.cuni.cz

Contact: Institute of Czech National Corpus (ucnk@ff.cuni.cz)

Gigaword Corpus

Both English and French gigaword corpora are a comprehensive archive of newswire text data in English that has been acquired over several years by the LDC.

Licence: Charles University owns a licence based on its LDC membership. For non-internal uses LDC User Agreement for Non-Members holds.

Languages: EN, FR

URL: https://catalog.ldc.upenn.edu/LDC2003T05

https://catalog.ldc.upenn.edu/LDC2011T10

Contact: LDC office (ldc@ldc.upenn.edu)

4 Conclusion

In this report we presented the first version of the toolkit for adaptation of the Khresmoi Translator to new languages. This includes a set of user-friendly scripts for training the statistical models for new languages to be used in the Khresmoi Translator. We also present a collection of data that can be used to train the models for English-Czech, English-German, English-French (the language pairs already supported by Khresmoi Translator) and also English-Hungarian, English-Polish, English-Spanish, and English-Swedish (languages to be added within KConnect).

5 References

[1] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst (2007). Moses: open source toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pages 177–180, Prague, Czech Republic.

[2] Aleš Tamchyna, Ondřej Dušek, Rudolf Rosa, and Pavel Pecina. MTMonkey: A Scalable Infrastructure for a Machine Translation Web Service. In The Prague Bulletin of Mathematical Linguistics, No. 100, pp. 31–40, 2013.