Masaryk University
Faculty of Informatics

English-Czech Dictionary for Kindle

Bachelor Thesis

Jakub Perháč

Brno, Spring 2015

Declaration

Hereby I declare, that this paper is my original authorial work, which I have worked out by my own. All sources, references and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Jakub Perháč

Advisor: Mgr. et Mgr. Vít Baisa

Acknowledgement

I would like to thank my supervisor, Mgr. et Mgr. Vít Baisa for his patience and helpful comments. I would also like to thank Mária Luptáková for proofreading and ideas, Lukáš Ručka and Ľubomír Remák for technical assistance and guidance.

Abstract

The aim of this bachelor work is to create a more efficient way of using the dictionary in users' Kindle e-book readers, providing them with English to Czech translations. The pre-installed monolingual Oxford Dictionary is often not sufficient, as it uses circular definitions, meaning that the term being defined is a part of the definition. Free alternatives are insufficient too, because of their limited vocabulary (i.e. containing only basic word forms). The customised dictionary is created from the uploaded e-book using constructed web-service. The script run by the web-page also force downloads the created dictionary to user's computer.

Keywords

Amazon Kindle, E-book, GNU/FDL Dictionary, English - Czech translations, Corpora

Contents

1 Introduction

Electronic readers and electronic books are a relatively new worldwide phenomenon. They allow us to obtain the information almost instantly at any moment and at any place if the internet access is available. Reading is more attractive and simple thanks to the accessibility of digital texts. Another quality contributing to the attractiveness of reading is the mobility of the electronic book equivalents. You do not have to take many (often heavy and bulky) printed versions, the digital ones will suffice. Also, e-readers are usually equipped with dictionaries, which is a great asset and many people tend to take the advantage of that.

Amazon's Kindle is a set of hand held devices dedicated to mediating the reading of the electronic books. These apparatuses have pre-installed Oxford English Dictionary, making them very powerful reading gadgets. However, the problem with the Oxford English Dictionary is that it sometimes contains vague or incomplete definitions that are not helpful at all. To meet the needs of non-native English speakers, who stumble across these obstacles, dictionaries with direct translations are required.

While there already is at least one free dictionary translating into Czech, its vocabulary is limited by the database of an open dictionary, therefore does not contain every form of the given word.

The aim of my thesis is to create a dictionary based on GNU/FDL dictionary, using English and Czech corpora. Specifically, it will be expanded by every word form used in the English corpus, making the resulting dictionary more extensive and therefore more viable. The web-page allowing people to generate their own customized dictionaries for individual e-books is also a part of this work. This web-page will be simple, as it is not the main focus of the bachelor work, however, it will provide the required service. The guaranteed functionality of this web-page is provided with the two tested web browsers, Google Chrome and Mozilla Firefox.

In the first part of the thesis, basic information about e-books, e-book readers, corpora and other issues are explained. Later, the existing alternatives, their pros and cons will be presented. After that, the importance of building a new dictionary and its added value will be analysed. Using (not only) these materials, the thesis will describe the process of creating a new dictionary and generating the customized ones. At the end of this work, the effectiveness of the new dictionary will be evaluated by testing the coverage of translated words on a few electronic books.

2 Background information

In this part of my thesis, a few concepts are introduced and the elementary information required for understanding the subject is given.

2.1 E-books

Electronic books (also called eBook, e-books, ...) are usually the electronic counterparts of the printed books, but it is not out of the ordinary, that there are some without a printed equivalent. They are meant to be read on the designated hand held devices called e-book readers.[1] However, we can read them on almost any electronic devise with controllable viewing screen (e.g. computer, smart-phone, tablet). The greatest advantage of using e-books is in its portability, dramatically decreased weight and volume of kept literature, simplified referencing due to electronic bookmarks or annotating pages.

Although fiction and non-fiction writings are common amongst e-books, technical texts are especially suited for this format. These texts can be searched (e.g. for documentation), copied (e.g. programming code examples) and processed significantly more effectively than their printed versions. That is the reason why technical paper books often contain CDs or other mediums.[2]

The first electronic book was created by Michael S. Hart. During his interview in 2002 he said: „We were just coming up on the American Bicentennial and they put faux parchment historical documents in with the groceries. So, as I fumbled through my backpack for something to eat, I found the US Declaration of Independence and had a lightbulb moment. I thought for a while to see if I could figure out anything I could do with the computer that would be more important than typing in the Declaration of Independence, something that would still be there 100 years later, but couldn't come up with anything, and so Project Gutenberg was born,"[**?**]

2.2 E-readers

The e-readers, also called the e-book readers, are hand-held portable devices designed mainly for reading digital e-books and periodicals. Unlike devices that are not dedicated to this purpose, these readers have increased readability (especially in the sunlight) and battery life (it is not unusual charging an e-reader after two months of reading). Also as opposed to the printed versions, e-readers are capable of containing hundreds of books with no measurable mass or weight.[4]

This is possible because of the electronic paper or the electronic ink, technology mimicking ordinary paper, allowing readability in a direct sunlight without any significant fading of the text. E-book reading devices use energy only when the shown screen is changed. The energy is used by the electronic ink. It consists of the magnetic particles, which are repositioned using the energy when the page is changed.
One of the first designated devices came out in 1992, when Sony Corporation introduced Data Discman. It was supposed to allow quick reference finding by searching pre-recorded disc.[5]

Amazon's Kindle readers belong amongst the most popular. Their set of devices is very well known and widely accepted, from Kindle (first generation that was released in 2007 and sold out within five and half hours[6]) to the newest Kindle Voyage.[7]
Company understood the importance of the books in the electronic format and in late 2009 introduced Kindle for the PC application, allowing people without e-readers to take the advantages of non-printed versions of books (purchased from Amazon's store) on computers running Windows operating systems (now available for Windows 8, 7, Vista and XP)[8]. In 2010 versions for Apple's Macintosh.[9]

Primary extension used by Amazon is Kindle's proprietary format AZW (which is similar to MOBI). This does not mean that Kindle is not capable of loading content in other various formats. Not only it allows users to upload a document of several formats (i.e. DOC, ZIP, HTML, MOBI, EPUB, RTF, TXT, PDF, KPF [10]), but the file conversion can be done via registered email address (converted file is com-

fortably downloaded to the e-book device through Wi-Fi). Another way of transferring data to Kindle is by connecting the device to a computer using USB cable.

2.3 Corpora

Apart from e-books and e-book readers, to meet the requirements and build a viable dictionary generator, basic understanding of corpora is necessary. They are the extensive collections of the lingual data. The incorporated texts were either written, resulting in text corpora, or recorded and their transcripts were made thereafter, creating the speech (also called spoken) corpora. There are both, multilingual and monolingual occurrences. After the textual data gathering, they are processed, purged of duplicities and anything that is not a part of the document's content. The refined corpus is parted into tokens (i.e. words, punctuation, not-words). Corpora are often indexed, for instance parts of speech are assigned.[11]

2.4 Dictionaries

We must not forget about a very significant and important feature e-book readers provide, which is dictionary. Amazon's Kindle is pre-equipped with the descriptive Oxford English Dictionary (OED). This dictionary is very useful and helpful for understanding (not only) technical texts. A very convenient way of getting a description of an unknown word is also worth mentioning. It is done simply by moving the cursor in front of the unknown word.
Although the importance and impact of this dictionary cannot be omitted, non-native English speakers can sometimes struggle with some obscure or vague descriptions, such as worker – somebody who works. Also, this can become a problem when a translation of an animal or plant is needed and the definition does not suffice. As an answer to this problem, dictionaries with local translations were created.

2.5 GNU/FDL Anglicko - Český slovník

The aim of GNU/FDL Anglicko - Český slovník project is to produce the open extensive English - Czech dictionary without using any other dictionary (printed or digital) that will be downloadable to a computer.[12] As a result of help from many volunteers, the considerable amount of translations has been added into the dictionary, creating a reasonable unit. In spite of that, this work lacks many word forms, therefore is not absolutely comprehensive and entirely applicable all the time.

This dictionary uses the ISO-8859-2 coding and has a very simple structure:
"English word [TAB] Czech word [TAB] comment [TAB] special comment [TAB] name of a person who added the word"

If more than one comment is required or necessary, they are separated with [SPACE]. These comments may indicate word classes (e.g. ":n" for noun, ":v" for verb, ...), sex (i.e. "[female]", "[male]") or area of used translation (e.g. "[astr.]" for astronomy). Also, it indicates whether the word is plural ("pl.") or singular (no comment is used).

Special comments serve as further explanations of the word or references to other translations.[13]

2.6 Calibre ebook management

Having a working and extensive dictionary is one thing, its ability to cooperate with e-books is the other. In order to access data from the given text, process it, build a dictionary and upload it in the e-book reader successfully, the format conversion is necessary. The generator based on this thesis works with MOBI and EPUB formats, as these are the most common amongst e-books.[14] The calibre ebook management is used to achieve that goal.

This free open-source application created by and for users of e-books has many features. The e-book conversion is one of them. It is capable of converting from and into a huge number of formats.[15] This conversion can be automatized, using the command line interface which is the part of the application. It even allows users to pass the arguments, incorporating them into the conversion process, such as transliterating Unicode characters into an ASCII representation, inserting meta-data and many others. What is more, it automatically detects the e-book structure and keeps it intact (e.g. chapters or Table of Contents).

3 Existing solutions

Now that the basic background information and fundamental knowledge needed for this project is explained, we can continue with revealing and analysing existing solutions. This section of the work discusses them, evaluates their pros and cons and compares the aforementioned.

3.1 The Amazon's ENGLISH-CZECH Dictionary With Transcriptions

One way of resolving the issue with OEDs obscure definitions is to buy the Amazon's ENGLISH-CZECH Dictionary With Transcriptions. This dictionary is very useful for Czech speakers as it is able to translate different word forms (e.g. reading, translated, etc.). It also can be set as a default Kindle dictionary, allowing the translations while reading. Furthermore it also contains transcriptions, so users can pronounce the words they are translating correctly.[16] However, not everybody agrees with buying the dictionary, as there is one already implemented in the e-reader.

Pros

- Capable of translating while reading
- Claims to have many different word forms
- Contains Transcriptions

Cons

- Not free

3.2 The integrated English-Czech dictionary by thomaskiranti

A free alternative to the Amazon's dictionary is a successful one based on the open GNU/FDL dictionary. It allows you to simply download it into the Kindle, set it as the primary dictionary and use it while reading other books just as the Amazon's one does.[17] Unfortunately, this dictionary often does not contain a word in a certain word form, therefore cannot translate it. For example, it contains the word autobiography but does not contain its derived word form, autobiographical.

Pros

- Capable of translating while reading
- Free

Cons

- Limited by the GNU/FDL database

3.3 Creating own dictionary

Another free way of getting an English to Czech dictionary is to create one. The Kindle allows you to import your own dictionary and set it as the primary one. Creating your own dictionary is not that difficult, there are plenty of instructions and how-to-s on the internet. However, in order to make a good dictionary with a sufficient vocabulary, a large dataset of translations is required.[18]

Pros

- Capable of translating while reading
- Created to suit your needs
- Free

Cons

- Time consuming set up
- Limited by the provided translation database

3.4 Comparison

For purposes of the English literature reading simplification, we are comparing three solutions of obtaining and using a dictionary.
First of them is the Amazon's unified and extensive one. Although powerful and helpful, the downcast of this dictionary is its price. However, for users who do not mind paying for the additional dictionary in their devices, this is probably the best option.
On the other hand, users who are not willing to buy a dictionary have two options, creating their own, or downloading one from the internet.
The easier and more acceptable way for majority of people is the latter alternative. They do not need any special set of skills nor do they need to be tech savvy. All it takes is the capacity to read with understanding and the ability to download a simple file into the e-reader.
More computer-advanced people, who are willing to spend their time learning something new and want to enjoy the results of it, should go for making their own dictionary, as it allows them to customize the resulting product the way they need.

4 Used technology

4.1 Python

Python is a high-level programming language developed under an OSI-approved open source license. That means it is freely distributable and usable, even for commercial use. It is general-purpose and very easy to learn thanks to its design, which emphasizes readability of the code. It also allows the code to be written in fewer lines than the same code in Java or C++.
It supports imperative, functional and object-oriented programming and has a large standard library.
The core philosophy of the language is summarized by the document "PEP 20 (The Zen of Python)"[24], which includes aphorisms such as:

- Beautiful is better than ugly
- Simple is better than complex
- Complex is better than complicated
- Readability counts

4.2 PHP

PHP is a scripting language designed for building web-sites, however, it is also used as a general-purpose programming language. This free software can be easily mixed with HTML code, which makes it highly favourable and widely spread. Its code needs to be processed by a PHP interpreter. This interpreter is usually installed on a web server that runs the web-pages using the interpreter.
Until 2014, no written standard existed, but there is ongoing work on creating a formal PHP specification[25].

5 Implementation

In this thesis, new English to Czech dictionary is created. It is based on the GNU/FDL Dictionary and benefiting from its open database. To avoid the limiting factor of the translation database that is incomplete, we are using the BNC corpus to our advantage, meaning, that every word form in the corpus is used to extend the original dictionary. Furthermore, this project includes also a website granting users a straightforward way of creating and downloading a customized dictionary for the specific e-book.
The added value of this project is the independence of the script on the specific corpora or dictionaries, making it an efficient base for further improvements. Thanks to this fact, this build also allows us to create other than English - Czech dictionaries.

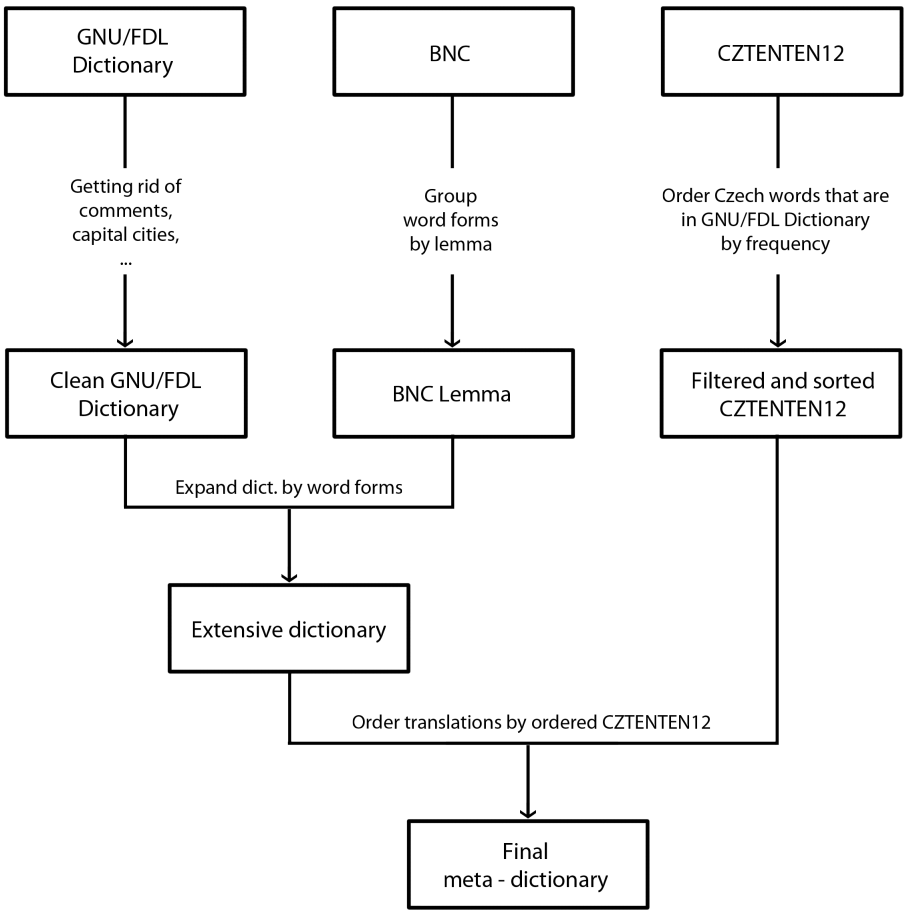
5.1 Overview

First of all, it is necessary to download the GNU/FDL Dictionary and modify it, meaning slight changes in its content and structure. Some adjustments in the BNC and the CZTENTEN12 corpora are needed too.
Secondly, merging with the BNC corpus is required, expanding the original dictionary on unused word forms.
The next step is sorting the translations by their frequency, which can be determined by the CZTENTEN12 corpus.
These parts together form the complete extensive English to Czech dictionary. What we need to do next is to compose a script, which allows us to take a book in an electronic format and return a personalized dictionary made for the e-book.
Creating a website that allows us to access the dictionary generator and download the result is the last step.

5.2 Creating a dictionary

In this section, making of the dictionary, expanding it by words from the BNC and ordering translations by frequency is explained (Figure 5.1).

Figure 5.1: Creating a meta - dictionary



5.2.1 Processing the GNU/FDL Dictionary

The dictionary is available in multiple formats and encodings. For our requirements, the utf8 encoded txt file seems to be the most practical. It supports special Czech (not only) characters[19], thus we start with downloading this one (Listing 5.1).

Listing 5.1: Downloading dictionary

```
import urllib.request
import shutil
url ='http://slovnik.zcu.cz/files/slovnik_data.txt'
with urllib.request.urlopen(url) as response,
    open('gnu-fdl.txt', 'wb') as out_file:
    shutil.copyfileobj(response, out_file)
```

The next step is to process a downloaded file. Starting with creating a list of non-desired translations (e.g. Capitals) and comments using the script that bans them (Listing 5.2).

Listing 5.2: Banning translations

```
defcorrect(words):
    if(words[0] ==words[1]):
        returnFalse
    elif(words[0] =='#'):
        returnFalse
    elif(words[1] ==''):
        returnFalse
    elif('capital of'inwords[1].lower()):
        returnFalse
    elif('hl. m'inwords[1].lower()):
        returnFalse
    elif('okres v'inwords[1].lower()):
        returnFalse
    ...
    else:
        returnTrue
```

Creating the ban list is followed by changing the structure of the dictionary, joining translations, simplifying the readability and usability of the current file (Listing 5.3). The form of a result is shown below: English word: translation1, translation2,... (e.g. cat:Kočka,Kočičí) .

Listing 5.3: Banning translations

```
defcleanDict():
    gnuFdl =dict()
    for line in sys.stdin:
        words =line.split('\t')
        for word in words:
            word =word.strip()
        if(len(words) > 1 and words[0] !='' and correct(words)):
            if words[0] not in gnuFdl:
                gnuFdl[words[0]] = set([words[1]])
            else:
                gnuFdl[words[0]].add(words[1])
    for word in gnuFdl:
        print(word +','+'{0}'.format(',').join(gnuFdl[word])))
```

5.2.2 Processing the BNC

In order to implement the BNC into the refined dictionary, it needs to have a certain desired form. Also, trimming redundant information is indispensable, as this significantly shortens the processing time. Only the first two rows of the corpus are taken from the input and the rest is omitted. This leads to the corpus having a format as displayed: Word ford [TAB] Lemma [-] shortcut for part of speech (e.g. attested[TAB]attest-v) This pattern is sufficient for the next step, compilation of the BNC. Searching for the word forms and joining them with lemmas for the easier parsing resulting in the structure similar to the adjusted GNU/FDL Dictionary:
Lemma [:] word form 1 [,] word form 2 [,] ... (e.g. card:cards,card)

5.2.3 Processing the CZTENTEN12

Complete dictionary's translations will be ordered by their frequency, therefore getting rid of the superfluous data in this corpus is necessary. In case of omitting or skipping this step, the runtime of the script generating meta-dictionary would be immense.
Cleaning is arranged by the code, which compares Czech word to translations in GNU/FDL Dictionary. If not found, it is not used in further development.
The remaining data are ordered by their frequency (determined by the frequency tag in the corpus) in the descending order, meaning that the most used words are located at the very top of the document and those not used so often are on the bottom. Also, to save some space and shorten the time needed for completing the sorted set, only the data with the frequency of 5 or higher were used.

5.2.4 Merging BNC and GNU/FDL Dictionary

The following part of the project answers the issue with deficient English vocabulary. Adding the BNC's word forms into the dictionary solves this problem by searching the English words, finding their lemma and joining the translation for every word form caught.
Example:

- Before merging
  - Cat:Kočka
  - Cats is not in the dictionary

- After merging
  - Cat:Kočka
  - Cats:Kočka

Although the translation for Cats is neither ideal nor precise, at least, it gives a hint or a preview of the correct translation and makes the understanding easier.

5.2.5 Sorting translated words

Dictionaries tend to have their translations ordered by a key. In this case, the key will be a frequency of the given translated words.
The frequency mentioned above is determined by the already processed CZTENTEN12 corpus.
The principle of the data ordering in the translations is pretty simple, but it takes a considerable amount of time to run the required script. The Czech corpus is parsed line by line, searching for the translations in every word of the dictionary. If the word from the corpus is contained in the current English word's translation, it is added into an ordered list. This list is assigned to the current English word.
Having the translations sorted is only a part of this section. The next step is to order the set of English words alphabetically. This can be done really simply, using the Python's built-in sorted() function.

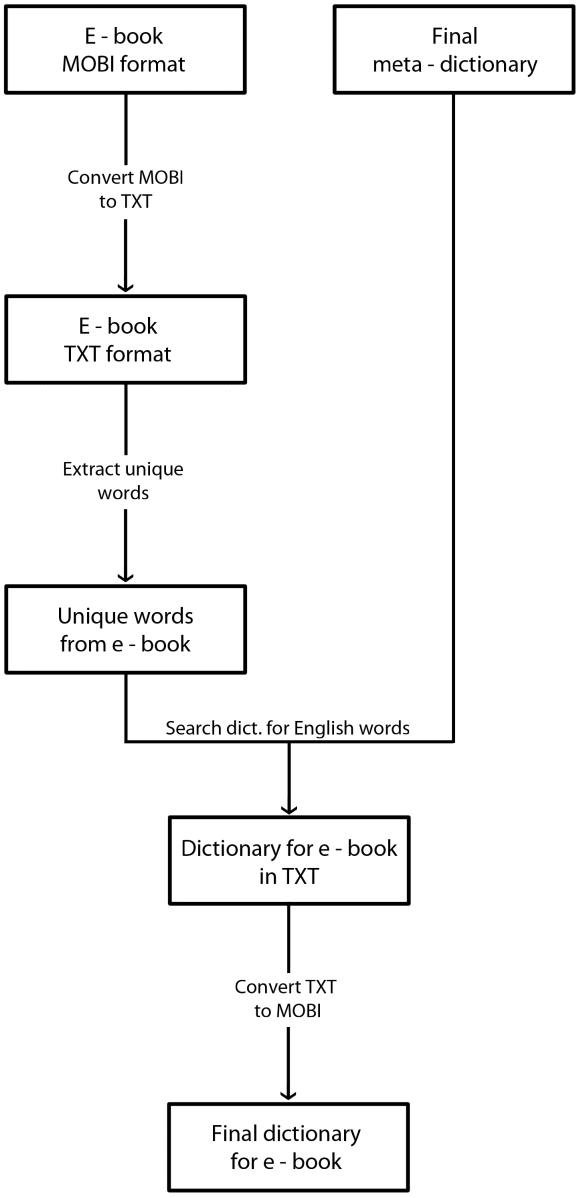
5.3 Customized dictionary

Amazon's Kindle devices have limited storage capacity with no possibility of expanding (except for the first generation)[20], so it is crucial for the dictionary to be as small as possible. A reasonable way of achieving this is by creating specialized and customized dictionaries, specifically, creating only one dictionary for a currently read book.
To accomplish that, we need to take the user's book, convert it into a text file making it accessible, then extract all words from it, find their translations and create a dictionary out of them. This dictionary needs to be converted to a Kindle supported format, MOBI (Figure 5.2). The free and open source program Calibre ebook management[15] is used for the automated converting from and into MOBI taking the advantage of the incorporated command line interface[21] (Listing 5.4).

Listing 5.4: Converting from and into MOBI

```
ebook-convert $(book) $(basename $(book)).txt
ebook-convert $(basename $(book))_dict.xhtml $(basename
    $(book))_dict.mobi
```

Figure 5.2: Creating a customized dictionary



5.3.1 Words extraction

For our purposes, we need every word from user's already converted e-book only once. Using of the Python's class set (unordered collection of unique elements)[22] accomplishes that (Listing 5.5).

Listing 5.5: Extracting words from plaintext

```
deftextReader():
    wordSet =set([])
    for line in sys.stdin:
        words =re.findall(r'[^\W\d]+',line)
        for word in words:
            wordSet.add(word.lower())
    for element in wordSet:
        print element.strip()
```

5.3.2 Creating the dictionary

With the help of the e-book's set of words, we are able to search the already created meta-dictionary for the translations and make a new specialized dictionary.
The last step is converting the result into MOBI format.

5.4 Web implementation

The main focus of this thesis was to create the project, which makes the construction of own dictionary possible. However, in order to allow accessing and using the aforementioned process of creating own dictionary to users, the best practice is creating a web service, therefore a simple web-page is necessary.

Firstly, user needs to upload an e-book to server, which runs the script creating a new dictionary. Secondly, the e-book must be processed by the above mentioned script and after that, the generated dictionary has to be passed to the given user. In the upload process, the basic security measures are implemented, allowing users to upload only

files with MOBI or EPUB extensions that are size of 5 MB or smaller (Listing 5.6).

Listing 5.6: Security measures

```
if(($ext=='mobi'||$ext=='epub')&&
    ($type=='application/x-mobipocket-ebook'||
    $type=='application/epub+zip'))
    {if($size<=$max_file_size) {
       ...
       } else {
           echo'File has to be 5MB or smaller';
       }
   } else {
       echo'File has to be mobi/epub';
   }
```

Multiple headers are set in the part of the code dedicated to the force download. One of these headers was modifying the type of the header of the downloaded file, specifically, replacing it, therefore making it corrupted and impossible to open. This was fixed by adding the FALSE parameter into the header, so it would not replace the original one, only add the second header[23] (Listing 5.7).

Listing 5.7: Content-Type header

```
header('Content-Type: application/octet-stream',FALSE);
```

6 Effectiveness of the dictionary

The number of the unique words in an e-book's plain-texts and in the generated dictionary are compared to measure the effectiveness and usefulness of this project. Using of the already created parts of the code combined with a few bash commands allows us to process a few books obtained from the Project Gutenberg and determine the coverage of the translations.
Firstly, the user's e-book is stripped and processed into the list of unique words. Secondly, the generated dictionary is searched for the occurrence of the words from the unique words list and total number of appeared translations is written down.
Similarly, GNU/FDL dictionary is searched and the result is noted.
The last step is the evaluation of the created and GNU/FDL dictionaries, their effectiveness based on the coverage of translations from randomly picked books.
Before analysing the books and testing them for translation appearances, the GNU/FDL Dictionary and our meta-dictionary were compared. While the GNU/FDL based one has almost eighty-five thousand unique English words with the associated translations, the one generated in this project contains over two hundred ninety-three thousand unique English words that can be translated.
The coverage of the generated dictionary and the GNU/FDL dictionary on three out of several tested books were following:

- Alice in Wonderland
  - Total number of unique words: 3012
  - Unique English words in GNU/FDL d.: 2793
  - Unique English words in generated d.: 2897

- The Adventures of Tom Sawyer
  - Total number of unique words: 7639
  - Unique English words in GNU/FDL d.: 2136
  - Unique English words in generated d.: 3302

- Ulysses
  - Total number of unique words: 29164
  - Unique English words in GNU/FDL d.: 2614
  - Unique English words in generated d.: 20067

Rest of the tested books' statistic is similar, meaning that the coverage of the new customized dictionary is better than the GNU/FDL Dictionary's coverage. This difference is ranging from 2% to 60% of the unique words contained in the e-book and the associated dictionary, depending on the tested text, while the most common value of the difference between the coverage was about 5%.

Based on the information gained from our testing, this bachelor work's dictionary is the best for more demanding texts, as simple and not very trying writings are already decently covered by the GNU/FDL based dictionary. However, even in these texts, there is a visible improvement in vocabulary capacity.

7 Conclusions

In this paper, we focused on building the custom dictionary generator designated to e-book readers, specifically to the Kindle. Foremost, basic background knowledge and information required to fully understand the issue of creating the Kindle dictionary generator was explained.
The main focus of the thesis is on the dictionary and the customized dictionary generator itself. The implementation process and single steps in it were thoroughly described, explicitly, processing corpora, the GNU/FDL dictionary, their merging and sorting, incorporating own e-book resulting in the customized dictionary and building the web-page.
The last part of this work consisted of analysing the effectiveness and viability of the aforementioned custom made dictionaries. Analyses were shown on three books downloaded from the Project Gutenberg and clearly implied, that the new dictionary has better coverage, thus is more applicable and usable.

The aim of this thesis was to create an effective dictionary generator, providing the English non-native speakers with the functional alternative to OED with the possibly better impact on reading English texts. After taking the results of testing into consideration, it is evident, that this bachelor work is a considerable option.

Also, thanks to the used techniques and created scripts, the dictionary generator does not depend on specific corpora, therefore it is possible to use this project as a platform for the future development of dictionaries in languages other than Czech and English.

Bibliography

[1] Oxford Dictionaries - Dictionary, Thesaurus & Grammar. [Online]. e-book: definition of e-book in Oxford dictionary(American English), Available: http://www.oxforddictionaries.com/us/definition/american_english/e-book.

[2] PCMag.com. [Online]. e-book Definiton from PC Magazine Encyclopedia, Available: http://www.pcmag.com/encyclopedia/term/42214/e-book

[3] The Guardian. [Online]. Michael Hart, inventor of the e-book, dies at 64, Available: http://www.theguardian.com/books/2011/sep/08/michael-hart-inventor-ebook-dies

[4] NYTimes.com. [Online]. When an E-Reader Is Loaded With Books, Does It Gain Weight, Available: http://www.nytimes.com/2011/10/25/science/25qna.html?_r=1

[5] Wikipedia, the free encyclopedia. [Online]. Data Discman, Available: http://en.wikipedia.org/wiki/Data_Discman

[6] Engadged. [Online]. Kindle sells out in 5.5 hours, Available: http://www.engadget.com/2007/11/21/kindle-sells-out-in-two-days/

[7] Business Wire. [Online]. Amazon.com Announces Fourth Quarter Sales up 42% to & 9.5 Billion, Available: http://www.businesswire.com/news/home/20100128006703/en/Amazon.com-Announces-Fourth-Quarter-Sales-42-9.5...#.VVf2Kd-jnQp

[8] TechHive. [Online]. Kindle for PC Released, Color Kindle Coming Soon?, Available: http://www.techhive.com/article/181810/Kindle_for_PC_Released_Color_Kindle_Coming_Soon.html

[9] Engadged. [Online]. Kindle for Mac now finally available, Available: http://www.engadget.com/2010/03/18/kindle-for-mac-now-finally-available/

[10] Amazon Kindle Direct Publishing. [Online]. Available: https://kdp.amazon.com/help?topicId=A2GF0UFHIYG9VQ

[11] Centrum zpracování přirozeného jazyka. [Online]. Jazykovy Korpus, Available: https://nlp.fi.muni.cz/web3/cs/JazykovyKorpus

[12] GNU/FDL Anglicko-Český slovník. [Online]. Uvod, Available: http://slovnik.zcu.cz/uvod.php

[13] GNU/FDL Anglicko-Český slovník. [Online]. Format, Available: http://slovnik.zcu.cz/format.php

[14] Digital Magazine Software. [Online]. What are the Differences Between .epub and .mobi?, Available: https://www.3dissue.com/what-are-the-differences-between-epub-and-mobi/

[15] calibre - E-book management. [Online]. About, Available: http://calibre-ebook.com/about

[16] Amazon.com. [Online]. ENGLISH-CZECH Dictionary With Transcriptions, Available: http://www.amazon.com/ENGLISH-CZECH-Dictionary-With-Transcriptions-Suponau-ebook/dp/B0076SLHLA

[17] E-knihy, čtečky e-knih, Amazon Kindle, PocketBook. [Online]. Anglicko český slovník pro Amazon Kindle – fičuryna týdne!!, Available: http://www.ebooky.cz/anglicko-cesky-slovnik-pro-amazon-kindle-ficuryna-tydne/

[18] MobileRead Forums. [Online]. How to create your own mobipocket dictionary for any language, Available: http://www.mobileread.com/forums/showthread.php?t=20480

[19] AIKEN Kutná Hora. [Online]. Proč používám UTF-8. Available: http://www.aiken.cz/article/proc-pouzivam-utf-8

[20] Amazon.com. [Online]. Kindle e-reader, vailable: http://www.amazon.com/gp/product/B00I15SB16/ref=topnav_storetab_kstore

[21] calibre - E-book management. [Online]. Command Line Interface, Available: http://manual.calibre-ebook.com/cli/cli-index.html

[22] Welcome to Python.org. [Online]. Built-in Types, Available: https://docs.python.org/2/library/stdtypes.html#set

[23] PHP: Hypertext Preprocessor. [Online]. PHP: header - Manual, Available: http://php.net/manual/en/function.header.php

[24] Python.org. [Online]. PEP20 – The Zen of Python, Available: https://www.python.org/dev/peps/pep-0020/

[25] ITworld. [Online]. PHP gets a formal specification, at last, Available: http://www.itworld.com/article/2697195/enterprise-software/php-gets-a-formal-specification--at-last.html

Attachments

Another part of this thesis are the electronic attachments stored in the IS MU archive

- Source codes of the program
- The project documentation
- Source codes and data used for the effectiveness testing
- Source code of the text part of the thesis
- Thesis in the PDF format