Title: collaborative Multi-regional Chinese Lexicon These project is the ideas of the primarily author, Jeremiah Chung, based on his experience working with The Bible Translator's Assistant(<http://thebibletranslatorsassistant.org/>) and The Linguist's Assistant(<http://www.linguistsassistant.com/>), SIL along with his subsequent work “A Proposal for Open Source Multilingual English-Chinese FLEX Lexicons for Biblical Literature” at the 47th International Conference of Sino-Tibetan Language and Linguistics(Oct., 2014, ,Kunming, China), This writing is based on the R Markdown langauge format with RStudio share in the repository Github Repostitory: <https://github.com/JChungYS/Collaborative-Multi-regional-Chinese-Glossary/> This proposal will be presented at Researching Collaborative Translation: An International Symposium, scheduled on 7-8 April 2016, hosted By the Centre for Translation, Hong Kong Baptist University. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

Demand for translation of global Chinese languages keeps increasing rapidly owing to globalization and localization through the Internet and other technologies. One of the challenges for satisfying this demand is to match the variations of the styles, vocabulary with the expectations of diverted readers and/or audience of Mainland China, Taiwan, Hong Kong and other overseas Chinese.

This paper is to propose a collaborative open source multi-regional Chinese lexicon based on git-hub platform as a tool to bridge the gap among these regional variations through natural language processing and different computational linguistics technology.

This paper includes a description of prototype of the tool, followed with an analysis of the strength, weakness, opportunities, and threats of this project. The scope of the project will start with the domain of Bible translation, which is the primary trade and subject of studies of the author over the last twenty years. There are significant translation studies and practices in such domain ever since the 19th century. This project can also be used as a model for similar lexicon projects, thanks to the thousands of languages that have already been translated for the Bible.

The conclusion will cover the potential implementation and the potential of the extended contribution and further research in areas such as:

1. Mentality and application of open source and open data in translation practice, policy, and theory,
2. Ethics of collaborative translation and translation tools,
3. Continuum of computer aid translation,
4. Natural language processing,
5. Machine learning and translation,
6. Methodology of scale up, startup open source translation project,
7. Relationship between translation and technology,
8. Globalization and Internationalization beyond European languages and postcolonial methodology in translation.

By the time of this presentation, the project might have already been implemented. In such case, the presentation may include the update and evaluation of the project.

# Bio

From 1995 to 2014 Jeremiah Chung served as Translation Advisor, Linguist, Training Coordinator, and Consultant for Applied Linguistics and Language Technology with SIL (Vanuatu) and SIL (East Asia).

He earned his Doctor of Ministry in the Bible Translation track at Gordon-Conwell Theological Seminary (USA), Graduate Diploma in Anthropological Linguistics (Australia), MCS (USA), BCS (HK) and Diploma in Applied Linguistics (Singapore). He is also the author of A descriptive grammar of Merei (Vanuatu) and a presenter at the international conferences of Bible Translation and Linguistics. Currently he is a consultant for natural language generation projects with The Bible Translator's Assistant.

# Introduction -- What is collaborative open source multi-regional Chinese lexicon?

## Collaboration

In the field of Bible Translation, the team and organization must consider the process and the quality of Bible translations because of the belief that the Biblical text is divinely inspired. There have been more interactions between different Bible Translation organizations and practitioners in last decades. One of the developments emerging from these interactions would be the new approach in Bible Translation with non-traditional practitioners, non-traditional methodology, and language technology, such as natural language processing. This focus seems to put more emphasis on the output/transformation, acceptability, and community or grass-root ideals in contrast to the mentality of traditional missionary Bible Translation. It also shifts the focus from the publication of scholarly work of Bible Translation as the end to the collaboration with grassroots users of the transformation as the never end process of Bible Translation with ever changing technology. Open Source and BT pattern in postcolonial context There are also postcolonial shift from missionary and scholarly focus projects based on colonial/imperial mentality to users’ community centered projects and involve non-traditional human resources, such as diaspora community and laity (Chung 2014). The traditional ownership of Bible Translation project is also being discussed and reflected along with the development of open source collaborate technology. In such a case contemporary Bible Translation practitioners realize the need of literacy program, endangered language preservation, orality for Biblical story along with other innovative strategies and tools. Recently, there have been new developments in Bible Translation methodology developed by Tim Jore and Distant Shore Media that are worth discussing. Jores' idea Bible Translation 3.0 and ideology of Christian Commons over copyright issues can be found and downloaded from the distantshores.org, where Bible Translation 3.0 is introduced as:

"The process of translating the Bible into other languages is returning to the global Church in the 3rd major wave of Bible translation. Characteristics of this paradigm shift include new strategies to reach every language in less time, as well as the removal of copyright restrictions from Bible translations in order to empower the Church to make full use of every technology for the building up of the body." ( <http://distantshores.org/> , visited Jan 4, 2016, Tim Jore)

## Emergent methodology about Open Source Collaboration

With cloud technology and crowd sourcing technology, translation can be done through the Internet with different degrees of participation of machine and larger team comparing with traditional Bible Translation team. With the trend of the publication of translation through some form of digital format, regardless of the differences of the degree of human/machine involvement, mapping of linguistics data between the source language and target language allows computing to enhance the effectiveness and efficiency in the translation process. The difficulty and process of translation depends on the difference between the source and target languages; however the human factor is largely minimized in translation owing to the increasing power of computer. In this paper, I will propose to start from collaborative open source dictionary of very closed or related variation or style of Chinese written language of Mainland China, Taiwan and Hong Kong. The dictionary is possible to be applied to different application and emergent approaches of Bible Translation.

## Translation as multi-disciplinary Technology

Whether translation is an art or science has been debated for a long time, but what interests me most is applying translation as technology, which means application of different disciplines to communicate across the boundaries of language and culture to bring a difference to the community and the world. Traditional translation involved many disciplines including communication theory, hermeneutics, cultural studies, different branches of linguistics, and knowledge of at least two languages. With the rapid development of technology and global changes, different disciplines such as engineering, technology and management, have also become the tools in the process of translations. In addition to all these disciplines, the knowledge of the discipline of the subject matter being translated, also plays an important role in translation. The primary interest in this proposal is about the collaborative crowd sourcing technology of digital lexicon for different degrees of computer aided translation.

As I mentioned before there are at least two languages, source text language and target text language. In Bible translation context there are two major source languages, Biblical Hebrew and Greek. Since there are more Greek New Testament translations being completed, in the COSMRD prototype project I’ll include Hebrew as the source language and Old Testament as the source text and English as the target languages. Based on that Hebrew-English dictionary and other English-target languages dictionary, we can develop different dictionaries of languages with Biblical vocabulary as a starting point for different purposes, e.g. translation, language teaching, and learning. What is Chinese?

Like most other modern Eurocentric translation studies, it is always ironic to me, as a native Chinese speaker, to talk about the Chinese language to other Chinese using English.

In my thesis of Engaging Chinese Diaspora, one key question is: what does “Chinese” mean? Interesting enough a lot of Chinese social studies are published in English. Chinese can be understood as a group of people according to their ethnicity, use of language, culture, origin of their homeland, nationality, etc. Chinese people and Chinese language have very subtle relation, but we’ll focus mostly on the Chinese language.

## The etymology of Chinese language

“The adjective Chinese is found in a few set phrases (Chinese copy, Chinese fire drill, Chinese money. and Chinese tour ) in which it represents inferiorityâimplying that something is less good, useful, or effective, or that it is not authentic. Although insulting a particular person or nationality may be unintentional, it is best to be aware that use of these terms is usually perceived as offensive to or by the Chinese.”

Qin Shi Huang unifiedying the writing system in his empire more than two thousand years ago. I was not able to trace the term “zhong wen” or “han yu” being used and the matching between the writing system and all different languages/dialects usinged in ancient China. In Guo Xi's The variation of Nomenclature of 'Chinese'(spoken) in Modern Chinese Society, the discussion is about spoken mandarin/pu tong hua(Guo 2007). That is a good starting point to understand the complexity of Chinese variation started from the nomenclature nowadays. The variation of Chinese can be a very complicated issues both historically and geographically. While Guo’s work is about spoken Mandarin/Chinese.(Putonghua), the focus on this paper is different contemporary written Chinese in the Mainland, Taiwan and Hong Kong. Written Chinese is quite different from spoken Chinese. However written Chinese and oralspoken Chinese could highly convertible as the voice processing technology is being quite developed. Thus, the lexicon data available can be a resource that allows for interaction between character encoding, spoken language with IPA(international phonetic alphabet) as phonetics tag and/or voice recognition, and written language of different variation of Chinese languages.

## Global Chinese vs. Regional Chinese

Language classification can be done with the ISO system such as Ethnologues. However Ethnologues primarily reflect spoken language information. An alternate resource for writing systems can be find in scriptsource.net. However, I fear that the not all Chinese writing system of different variations are documented for the purpose of natural language processing. The difference and relationship between writing system and the spoken language can still be very complicated, such as Chinese characters being used in Japan and Korea.

There are different variations of Chinese language along with different political implications. In general mandarin/putonghua/guoyu is considered as an international/global spoken Chinese and global written Chinese reflecting mandarin, it is not sure whether there is an agreed standard writing system matching all Chinese variations and dialects. The historical legacy Chinese encoding (Big 5 and Mainland's encoding) for computer and digital word processing is a good example reflecting the process of developing a common standard (unicode) as a unifying set of global Chinese characters. I fear there is not yet a set of global Chinese vocabulary with consensus by all Chinese. The Chinese Union Version Bible was a good attempt, at least in the domain of Christian churches as in the title \* A Study of the Phenomenon of Authoritativeness in the Chinese Translations of the Protestant Bible\* by Ruo Yu Zhuang(Zhuang 2000).

## What does it mean the regional diversity Chinese

### Mainland, Hong Kong, Taipei

Searching on the HK public library catalogue of “China, Hong Kong, Taiwan, “ yields a result with 290+ entries, but most of them are about economic, financial, political, and social issues but not many are about literature, language, and comparative studies in particular.

Searching on Amazon in English phrase, "China, Hong Kong, Taiwan" ends up with many entries. The first eight categories are about history, movies and political sciences. However, there were only 3 entries when I add the word "languages."

The publications and research projects of Chan Ho Yan Clara of Hong Kong City University are worth noticed, namely,

Books:

Chan, Clara Ho-yan (In-press).《法律翻譯系列：兩岸三地合約法主要詞彙》(Legal Translation Series: Key Terms in Contract Law of Hong Kong, Mainland China and Taiwan). Taipei: Bookman Books Co. Ltd.

Chan, Clara Ho-yan (2015).《法律翻譯系列：兩岸三地侵權法主要詞彙》(Legal Translation Series: Key Terms in Tort Law of Hong Kong, Mainland China and Taiwan). Hong Kong: City University of Hong Kong Press. (978-962-937-245-3).

Chan, Clara Ho-yan (2014).《法律翻譯系列：兩岸三地合約法主要詞彙》(Legal Translation Series: Key Terms in Contract Law of Hong Kong, Mainland China and Taiwan). Hong Kong: City University of Hong Kong Press, pp. 282 (ISBN 978-962-937-241-5).

Research Projects:

“The Intellectual Property Law Terminology of Mainland China, Taiwan and Hong Kong: Translation Strategies and Regional Variation”, Strategic Research Grant, CityU.

Other two books worth to notice is:

中国大陆, 台湾, 香港行政公文纲要, 张佐邦, 中国档案出版社, 1998.

From the searching of baidu.com and books.google.com, there is a translated work with title, China, Hong Kong, Taiwan, by Willem van Kemenad. The Chinese version published in 1998.

Again this together with other books with Mainland, Hong Kong, Taiwan is more about economic, social, and political aspects, but not much about communication and language. However, there is a comparative studies of Chinese literature in Taiwan and Hong Kong included in mainland China’s high school Chinese curriculum in the following article.

王驿钢. 2014. 高中语文教材中台港作家作品编选比较研究 作者出版源, 河南大学.

One interesting point of observation is the order of China, Taiwan, Hong Kong in the title of the above articles. As in academic writing of this paper, the order of these three are just following convention without any political preference or agenda.

## global verses regional Chinese with internet

There are cultural clash and conflict within the global Chinese sub-cultures and sub-languages. Outside the global Chinese sphere, non-Chinese and Chinese alike may underestimate the diversity of global Chinese. As I am using Chinese to browse the Internet, I found regional Chinese of China, Taiwan, Hong Kong Chinese is not just between simplified and traditional characters issues, the setting of locale, region and internet protocol address. How can the server be aware of my regional variation and my sub-culture category? My vocabulary and the choice of my languages and encoding system can be some clue. That can be another win-win situation or contribution from a multi-regional variation dictionary.

I fear that the importance of the comparative linguistics studies and inter-lingual communication within the Global Chinese is under-estimated. Besides the work mentioned above, there can be other literature and research about comparative linguistics or terminology of different regional Chinese in different domains that the author is not aware of. COSMRD with Biblical terminology can be a starting point.

# Methodology

The outcome of transformation/feedback based on translation can hardly be reduplicated with all different contexts. Thus, reproducible research methodology is supposed to fill the gap between nothing and reduplicate research as suggested in Roger Peng’s evidence based reproducible research method in Report Writing for Data Science in R, 2015. According to Roger Pengs, “the prerequisites of reproducible research are:

1. Availability of analytic Data
2. Availability of analytic Code
3. Documentation of code and data
4. Standard means of distribution” (Pengs：2015，20.1/142 )

COSMRD started as analytic data with two fold of downstream and upstream of two group of application or analytic code. The first group is to elicit and get the data and COSMRD as the documentation of code and data with standard means of distribution through github. The COSMRD will go through another set of analytic code of natural language processing to produce the translation and also distributed through github or digital reader/evaluation application. As it is not necessary to have timeframe for the translation projects and its revision for open source development contributed by volunteers, the cycle and the interaction between the data and human can keep going on and the data and the translated outcome can be continuously evaluated and refined.

## Code/Application and Data Interaction

Application code and data are like the debate on chicken and egg, which come first? According to Christian belief, human beings are created, and human(the hardware) created prior than human's words(the software) and the knowledge, skills, attitudes. However, it can be controversial at least theologically about God creating light and the world with God's Word prior to human with words. Thus, word or data could exist before human. Assume both data, code and hardware coexist, data and codes keep evolving. One of key areas we need to tackle is about how data being interpreted or computed by the codes. The codes and data are evolving to the status as today. We need to think about the matching of the format and/or patterns of data in order to produce desired outcome in an effective way, and the desired outcome can be extended to different applications and domains with innovation.

The primarily desired outcome of this project is to have a list of reference of different regional written Chinese terminology in different domains(e.g. Biblical term for Bible Translation, Linguistics, Translation Studies) with the context of machine aided translation for Bible Translation and providing resources of Bible Studies, Bible Translation and Scriptural Engagement for global Chinese users. One of the uniqueness about Biblical terminology is that the text of the Bible has been translated into hundreds of languages, quite likely the literature with the largest number of translations as a unique group of multi-lingual corpus.

## Production/Translation/Editing Tools

Let start think from the analytic code backwards to the data. There are different application from generic word processor e.g. notepad++, vi to some complicated natural language processing application, e.g. The Bible Translators Assistant, Google Translate, etc.

1. Scriptural Translation/Editing (Biblical Terminology, build in)

In general speaking scripture editing tools are different from generic word processer in its way of handling scriptural reference and providing other Biblical texts; different versions of Bible and other tools and resources for translation. In SIL and Bible Society linguistics and translation used to be two separate but related disciplines and software developers develop linguistics software to manage language data and use word processer for editing translated text.

Bible Translation software and references data base has been developing and adding different features, such as, reference materials, key-term list, language data management, text comparisons, interlinearization, project management, collaboration with team, send/receive through internet etc. over time. At the same time linguistic fieldwork software for managing and analyzing linguistics data want to integrate/connect with Bible Translation software. And there have been attempts of different computer aided translation.

-Paratext

“Paratext and related tools are a collection of software programs for Windows and Linux developed jointly by the United Bible Societies and SIL International which allow you to input, edit, check, and publish a translation of the Scriptures, based on the original texts (Greek, Hebrew), and modeled on versions in major languages. A growing body of scripture translation text in the USFM format is the basic unit around which all of the Paratext tools, utilities, and companion programs operate.”(Visited Dec 31,2015)

One of the concerns is the issue of releasing the software as open source has been suggested but has not yet resolved.

-Translator Workplace and BART

Translator Workplace and Biblical Analysis and Research Tools are proprietary Bible translation resources for practitioners who belong to or associate with project belong to some Bible Translation organizations. Some owners of the resources generously grant the exegetical and Biblical studies resources for the goal of Bible translation.

-Translation Studio

“translationStudio (tS) is a free utility for translating the Bible and biblical content into any language. The purpose of translationStudio is to make Bible translation accessible to the global church. translationStudio is fully functional without an internet connection and enables collaboration in the translation process. The app includes numerous just in time translation helps that explain keywords, difficult to translate phrases, and general translation concepts. When translations are completed, the app enables uploading and publishing through the unfoldingWord platform.” (<https://unfoldingword.org/translationstudio/>)

-Adapt it(For machine adaptation of scriptural text between related languages)

“Adapt It is a free open source tool for quickly translating between related languages. We seek to create, as a community, an adaptation tool that will run on all major platforms and facilitate translation of USFM, and collaboration between other translation tools such as UBS Paratext and Bibledit.” (<http://adapt-it.org/about-adapt-it/>, visited Dec 31, 2015)

-Toolbox/Shoebox(1980s-2015)

“Toolbox is a data management and analysis tool for field linguists. It is especially useful for maintaining lexical data, and for parsing and interlinearizing text, but it can be used to manage virtually any kind of data. If you have experience with the Field Linguist's Shoebox, you can think of Toolbox as an enhanced version of Shoebox with a new name. It is fully compatible with Shoebox. You can install Toolbox and use it just as you used Shoebox, with virtually no difference. Because Toolbox is an enhanced version of Shoebox, it is sometimes referred to as Shoebox/Toolbox. Toolbox is a text-oriented database management system with added functionality designed to meet the needs of a field linguist. The underlying dbms offers full user flexibility in the design of any type of database. But for ease of use, the Toolbox package includes prepared database definitions for a typical dictionary and text corpus. The Toolbox database management system offers powerful functionality like customized sorting, multiple views of the same database, browse view to show data in tabular form, and filtering to show subsets of a database. It can handle any number of scripts in the same database. Each script has its own font and sorting characteristics. While Unicode is preferred, Toolbox can handle scripts in most legacy encoding systems. Toolbox also has powerful linguistic functionality. It includes a morphological parser that can handle almost all types of morphophonemic processes. It has a word formula component that allows the linguist to describe all the possible affix patterns that occur in words. It has a user-definable interlinear text generation system which uses the morphological parser and lexicon to generate annotated text. Interlinear text can be exported in a form suitable for use in linguistic papers. Toolbox has export capabilities that can be used to produce a publishable dictionary from a dictionary database. …”( <http://www-01.sil.org/computing/toolbox> visited Dec 31, 2015) While toolbox is a relative light-weight tools primarily handling lexical data, it also provided function of adaptation between languages.

-Fieldwork Linguistics Explorer(FLEX)

“FLEx (FieldWorks Language Explorer) is a feature-rich program for dictionary compilation and text analysis. FLEx supports mutiple users colloborating on one project via a Send/Receive feature, which includes intelligent merging of work. cultural observations and morphological grammar testing. A command in FLEx (in the File menu) to start up RAMP and prepare an upload package of a FieldWorks backup file for submission to the REAP repository. FLEx works seamlessly on the Windows and Linux platforms allowing users on both platforms to collaborate ….” (<http://www.sil.org/resources/software_fonts/flex> visited Dec 31st 2015) Most of these tools are developed by SIL or its partner organization and posted in Lingtransoft. However, the translation software focuses on Bible Translation and seldom is made known to the academic community. However, with Nida Schools being associated with the American Bible Society, there seems have been more interactions between the Bible Translation Practitioners and academic circles in translational studies.

-The Bible Translators Assitant(TBTA)

“The Bible Translator's Assistant Inc. is ... a group of computational linguists and Bible exegetes who have developed software which enables missionaries and linguists to translate the entire Bible into a new language in a fraction of the time required by manual translation.” TBTA is also the name of the software to generate natural language text for Bible Translation based on the data of lexicon, transforming rules and ontology.

### Lexical data collection Tools and Application

Lexical data can be collected by words based method and text based method or simply getting existing vocab list and using tools to convert them to the needed format.

#### Words Based method

-Rapid Word Collection with WeSay

“The Rapid Word Collection method aims to revolutionize the task of collecting words by using a systematic method to capture these words in a workshop organized in the language community. Rather than the default language worker’s process of collecting words over a period of years and then publishing a work containing 5,000 words or so, RWC workshops consistently achieve a total of 10,000 or more raw entries during a brief two-week period. When compiled in lexical format, this will result in approximately two thirds of that amount as unique lexical entries with one or more senses. The only other method that comes close to the effectiveness of Rapid Word Collection is the text corpus method--the highly technical process of gleaning words from thousands of vernacular texts, which are in short supply for most languages which have only recently begun to be expressed in written form. The text corpus method requires a computer-savvy linguist, while the Rapid Word Collection method has been used in over a hundred languages by previously untrained native speakers.

See more at: <http://rapidwords.net/#sthash.DYsOLpxL.dpuf> and <http://rapidwords.net/”> We Say is a handy tool to collect data based on semantics and a list of question. However, it is important to have the list of questions in the language that the team can understand and with some field work technical support.

### Text corpus method

#### Multi-lingual corpus

-Paratext interlinearized function

Paratext based on a source text and a target text, it can generate an interlinearized text based on statistic. The interlinezied text can be extract as vocab pairs as a prototype dictionary entry. It can be very helpful for language have only Biblical literature. Howeve, I believe that further research and evaluations of the effectiveness is needed. The effectiveness depends on the styles of both texts, we expect the outcome will be more reliable with literal translation with large size of texts.

#### Monolingual corpus

-Using machine/online translate tools,e.g. google.translate or bing translate:

1. Generate list from a corpus e.g. a list of Union version: Taiwan, Hong Kong, Mainland.
2. Using google.translate or other translation tools, may need to cut down to smaller chunk
3. Clear up and Compile the result

### Lexical data format

The format of lexical data depends on collecting application and the analytic or translation application.

Possible format:

-text

One of the simplest format can be a text file with coma, tab or other symbols as separator, e.g. csv, which can be easily export to spreadsheet format

-standard format

Toolbox’s standard format, which can be easily exported to more complicated software like Fieldwork Linguist’s Explorer with LIFT format.

-LIFT

" LIFT (Lexicon Interchange FormaT) is an XML format for storing lexical information, as used in the creation of dictionaries. It's not necessarilythe format for your lexicon. That can be tied to whatever program you're using. But LIFT allows you to move that data between programs (hence the term 'interchange'). LIFT is also a decent archiving option. Not because it will be around in 50 years, but because people will still be able to read it with any text editor and easily make use of it, even then. (You think that's true of your non-SOLID Standard Format file? We should have a chat.)

LIFT has been designed to have a long life but also to be relatively easy to convert to and from existing lexicon formats, particularly Multi-Dictionary Formatter (MDF) and FieldWorks Language Explorer." (<https://code.google.com/p/lift-standard/>)

-Database format for The Bible Translators Assistant

TBTA may have quite complicated semantic tags such as substantive, mental predicates, AEM(action, event, movement), Existence, etc. eight-teen tags as in the list Natural Semantic Primitives (Goddard 1998:58)

#### Principle of choice for data format

The principle for the choice of data format from simple to complicated depends on:

1. the application of collecting the materials and the application managing the data
2. need of the application to analyses and/or translate
3. formats should be easily convertible

## Distribution of COSMRD

We need both lexicons and humans to produce better quality translations, proving that we need lexicons along with other data and tools. Listed below are several resources that may be helpful:

Repository

We need to have a server to keep all software, interface and data in open source license for collaboration.

GitHub

“GitHub”, which can allow storing web-site, wiki, script of different computer languages, text data, files of different format... as a version control and content management system. “GitHub” has been widely used in the software development industry with very strong open source and resources sharing core value. Repository is free for public usage. "GitHub is a Web-based Git repository hosting service. It offers all of the distributed revision control andsource code management (SCM) functionality of Git as well as adding its own features. Unlike Git, which is strictly a command-line tool, GitHub provides a Web-based graphical interface and desktop as well as mobile integration. It also provides access control and several collaboration features such as bug tracking,feature requests, task management, and wikis for every project. GitHub offers both plans for private repositories and free accounts, which are usually used to host open-source software projects. As of 2015, GitHub reports having over 11 million users and over 29.4 million repositories, making it the largest host of source code in the world." (Wikipedia, visit Dec 31, 2015)

-Pros:

1.Version Control System - keeps track of change highly portable with different languages highly compatible with web-based different platform.

1. Free for public repository.
2. Open source collaboration mentality behind creation.
3. Highly flexible for potential development in the future.
4. Maintained by teams of collaborators led by leaders with good reputation and track record.
5. Flexible in different formats of data

-Cons:

Learning curve can be a bit step.

Interface is primarily in English.

Repository of Language Depot: <http://public.languagedepot.org/> FLEX data in LIFT format “This site is provided as a service to language communities by the Language Software Group of the Linguistics Institute at Payap University, Thailand. It is primarily used by teams collaborating via Chorus-enabled software applications, including WeSay, FLEx), OurWord, & OneStory Editor. For dictionaries, two kinds of LanguageDepot project are available, choose one or both of them as necessary. FLEx Database: This kind of project is for synchronizing all of the data in FLEx, only FLEx can synchronise with it. LIFT XML: This kind of project can is for synchronizing only the dictionary data, both WeSay and FLEx can synchronise with it.” (visited Dec 31, 2015)

-Pros:

1. Privacy
2. Data Consistency Conforming with LIFT XML.

-Cons:

1. Proprietary
2. Only allow certain format LIFT for certain set of softwares/application
3. LIFT is complicated other public resources for Bible translation: Door43.org

### Criteria for Choice of Tools

#### Choice of lexical management/distribution tools

Consider the format first or the tool first

Format - SFM(Multilingual-format) LIFT

What formats for the lexical data what available and distinct features What kind of tag?

Issues relate with localization and tag for region in Genre, Style, localization for zh\_prc, zh\_hk, zh\_tw

1. Digital
2. Portable
3. User friendly
4. Import/export or convert to different format
5. Chinese interface?? Especially converted between Chinese
6. Allow bulk editing
7. Allow collaboration and merging data
8. Allow version control system
9. Allow extended tag for semantic representation
10. Allow different degree of complexity for different users

### Defining the final product

We also need to define the final product as:

1. the application making use of and evaluate the outcome of the COSMRCD,
2. the application that stores the Lexical data,
3. an application that mine or collect data and feedback.

There are a number of issues we need to consider.

-What should be the role of human practitioners, from software developers to monolingual vernacular language users?

-What skill sets are needed for all these pratitioners?

# SWOT

The following are some generic analysis of the strength, weakness , opportunity and threat for COSMRL.

## Strength

Github allow Collaboration

Involving global human resources in some innovative ways Using English or other major languages to bridge the gap between different languages.

Extension for different dialects, and different minority languages work along with language technology development - using language technology but engaging Chinese community for development and translation making language data available is the first step to making other literature available for community development and providing option for labor workers change into knowledge workers in the community

This project can benefit Chinese and others through better communication among the Chinese and also other cultures

## Weakness

The gap between the technical community and the non-technical community will grow even greater.

Limitations of digital text are not yet in audio format. Until we develop better voice recognition

Modern Babel Tower - There are dangers and severe negative consequence, if we are powerfully using languages and collaboration with evil motivation and narrow nationalism or ethnocentric mentality.

## Opportunity

Free public platform available for collaboration.

Extension for different Source languages: Biblical Languages Based, English Languages Based,

Allow more innovation for using their language for language technology and translations.

Extension for different domains

Domain Specific Consideration beyond Biblical material: Semantic Domain, Linguistics Domain, Translation Studies Domain, With transition of different languages

Start with Hebrew-English wordlist and merge with different language wordlist, e.g. English-Chinese wordlist, Chinese-minority language wordlist

It is important to create something from nothing for the template that volunteers and general public can work on. We can start with multi strategy and with collaboration with different degree of complexity and involvement of machine and different group of human resources .

## Threat

Need of the awareness of complexity and variation between Chinese dialect in the sense of computer linguistics and applied linguistics and at the same time value the historical and comparative studies for reconstruction of the protolanguage and history.

Too many technical details and software involve decisions of cooperation. Domain specific to reduce the size of the lexicons.

Typological specific to reduce the number of tags, hence the sized of the lexicons.

Disturbance of the economy and ecosystem of the current translation industry and Bible Translation industry.

# Conclusion

Tools and data are already available, the field of multi-disciplinary Bible Translation has a relative long history and lots of resources. Translation studies in Asia still tend to be very Eurocentric and bounded by departmentalism/sectionalism. Open source and collaboration in translation practice and translation studies can not only enhance the development of industry of translation and research on the subject matter of translation studies. Translation's nature of bridging gap between languages to allow all different academic disciplines to interact across the boundary of language is a win-win situation in the global context. Issues about Chinese and its variation, unity or diversity, in global Chinese context. Open Source Collaboration Technology and mentality in translation and different disciplines. Open Sources Collaboration mentality among Global Chinese, Chinese of different regions and Chinese diaspora, Chinese Christians Community with concerns on copyright of Bible translation and sharing resources in global context. Helpful for comparative linguistics, historical and social linguistics about the change of Chinese. If we do not tag or mark the difference now, how will this affect the difficultly and validity of the data in the future research? How will the tools and technology affect the form and the use of Chinese in the future? It would be good to take a snapshot now and compare it with the development of the variations in the future. For a number of minority in China, Chinese is the Lingua Franca and languages connecting them with the outside world. Translate more English into Chinese, and it will allow more resources to become available for minority languages. This knowledge and resource can be a significant help for the economy and community in the future of China, alleviating the living standard of the poor and the minorities in the rural areas. Forming communities of Amateur linguists, programmers, and translators to make more valuable resources to available to the general public in affordable ways. Implication of more collaboration involvement of machine and amateurs, we need to have good strategy to facilitate the change without hurting any of the existing ethical translation practitioners -- new generations of translators will be linguists and programmers that instruct the machine to interact with humans on a whole new level of competency.

## Issues/Areas for further research

The conclusion will cover the potential implementation and the potential of the extended contribution and further research in areas such as:

1. Mentality and application of open source and open data

In translation practice, policy, and theory, with modern technology of social networking, communication and computation, global connection and collaboration research and business are already possible; one of the barriers for the global collaboration is the language barriers. Translation is the technology and knowledge needed to bridge the gap. Sharing resources, knowledge, skills and technology in the open source public domain will certainly empower laity and general public to be engaged in global society.

How does this affect the ecology of global translation industry and the current translation practitioners and scholars/educators in the field of translation?

I think we need to handle the transition well and share a vision for the future development of cross cultural and interlingual translation with more aid from the computer and technology and collaboration of the communities and general public.

1. Ethics of collaborative translation and translation tools

In collaboration projects, how can people be accountable for machine aid translation?

For certain portion of the books done by different people, how can we keep the consistency?

What would be the ownership of a certain translation for open source projects?

What would be the legality issues and accountability for the owner of a translation projects?

How can we clarify and communicate our position in the disclaimer and decide what kind of open source or public licensed being used?

1. Continuum of computer aid translation

Based on different translation project, requirement of the commissioner, market and resources available, there is no one translation theory or approach for all translation need. There should be a wide range of translation approaches in the line of continuum with different degree of financial involvement and volunteer's engagement, different degree of involvement of machine aided process with different disciplines, different requirement of professional competency and/or training.

However, we should not assume that volunteers are less professional nor the qualities of machine aided translation will not as good as human translation. Comparing and evaluation of different translation quality needs to be done in a fair way. Depending on the outcome and the accomplishment according to the purpose of the translation projects, rather than who is doing the Bible translation. As for the data of the lexicon, syntax and corpus accumulating, we should make good use of different strategy with these resources to improve the quality and outcome of translations.

1. Natural language processing Knowledge based on information

Information based on data how can we make good use of linguistics data collecting and mining technology to develop knowledge base and the structure in order to generate information and knowledge especially cross the language boundary? What kind of languages mapping required? What should be the format and structure as a common Meta Lingual Franca to communicate between machine and human? What should be the tags and semantic representation for the lexicon, ontology, the source text and target text?

1. Methodology of scale up, startup open source translation project

The project can start up with relative low cost as an open source projects. Adding and checking tag and more meta-lingual information for lexicon items can be a lot of work, even we deploy automation through data mining or machine learning.

How about the same methodology applying to translation projects and as collaboration projects along with collaboration of the lexicon projects, interacting with each other’s?

1. Relationship between translation and technology

In what way Bible translation studies is different from general translation studies? How do we define translation? Why or why not translation is a kind of technology or language technology?

1. Globalization and internationalization beyond European languages and postcolonial methodology in translation

With the tools and translation technology bridging up the gap between the Chinese but non-English speaking community will enhance the global Chinese development. No one knows the future global Chinese, mainland Chinese could go divergent or convergent.

1. Application in Translation

This project can apply to related language adaptation, natural language generation and/or machine learning with relevant tagged lexicon format and content. Collaborate with the right upstream and downstream application can make the tools very powerful.

1. Allow change in a slow pace for the current translation practitioners to change

In Chinese Corpus, is it necessary to add the tag of different styles or variation, allow filtering and better data mining effectiveness? Allow different variation of languages is it a good thing or bad thing, depends on what context and language policy/planning?

# Bibliography

Chan, Clara Ho-yan (In-press).*《法律翻譯系列：兩岸三地合約法主要詞彙》(Legal Translation Series: Key Terms in Contract Law of Hong Kong, Mainland China and Taiwan)*. Taipei: Bookman Books Co. Ltd.

Chan, Clara Ho-yan. 2014.*《法律翻譯系列：兩岸三地合約法主要詞彙》(Legal Translation Series: Key Terms in Contract Law of Hong Kong, Mainland China and Taiwan)*. Hong Kong: City University of Hong Kong Press, pp. 282 (ISBN 978-962-937-241-5).

Chan, Clara Ho-yan. 2015.*《法律翻譯系列：兩岸三地侵權法主要詞彙》(Legal Translation Series: Key Terms in Tort Law of Hong Kong, Mainland China and Taiwan)*. Hong Kong: City University of Hong Kong Press. (978-962-937-245-3).

Chung, Jeremiah. 2014. *“A Proposal for Open Source Multilingual English-Chinese FLEX Lexicons for Biblical Literature”*. Power point slides presented at at the 47th International Conference of Sino-Tibetan Language and Linguistics, Kunming, China, October, 17-19.

Chung, Y. S. Anthony(Jeremiah). 2014. *“Engaging Chinese Diaspora Laity in Bible Translation”* DMin. Thesis, Gordon-Conwell Theological Seminary.

Jore, Tim. 2015. *Bible Translation 3.0* Accessed January 15, 2016. <https://unfoldingword.org/assets/docs/whitepapers/Bible%20Translation%203.0-rev%208.pdf>.

Jore, Tim. 2015. *The Christian Commons*. Accessed January 15, 2016. <https://unfoldingword.org/tcc/>.

Peng, Roger D. 2015. *Report Writing for Data Science in R*, LeanPub. Accessed January 15, 2015. <https://leanpub.com/reportwriting>

Van Kemenad, Willem. 1997. *China, Hong Kong, Taiwan*, New York:Knopf.

王驿钢. 2014. *高中语文教材中台港作家作品编选比较研究*, 河南大学.

张佐邦. 1998. *中国大陆, 台湾, 香港行政公文纲要, 北京* , 中国档案出版社.

莊柔玉.2008. *基督教聖經中文譯本權威現象研究*，香港：國際聖經協會.

郭熙. 2007. *现代华人社会中称说'汉语'方式多样性的再考察*， 南开语言学刊2007，第一期.