Building a Collaborative Workspace for Lexicography Works in Indonesia

Totok Suhardijanto[1], Arawinda Dinakaramani[2]

[1] Faculty of Humanities, University of Indonesia, Depok, East Java, Indonesia
[2] Faculty of Computer Sciences, Depok, East Java, Indonesia
E-mail: totok.suhardijanto@ui.ac.id, arawinda.dinakaramani@ui.ac.id

Abstract

This paper presents our attempt to develop a dictionary writing system for lexicographers in Indonesia. However, it does not mean that our work is only well-fitted for Indonesian languages. We developed this system from scratch to meet the basic need of lexicographers in Indonesia who are scattered in many local cities and prefer working in a team. A system which is designed and developed to meet our own demands is more easily adjusted than other existing systems. For this reason, we decided to develop this system rather than using existing ones.

In general, like other interactive lexicon viewing and editing applications, our system also provides hyperlinks for entries, category views, dictionary reversal, search engine, and export tools. However, our system is different to some extent. It is developed in a shared workspace concept to deal with lexicographers with geographic obstacles like in Indonesia. The system also comes with a corpus tool which allows users to create their own corpus. Users can store and access their language database from different locations. The corpus tool enables users to do corpus analysis and manipulation. Some major languages, such as Malay, Javanese, and Sundanese, are provided with grammatical annotation services. So, based on language corpora, users can perform lexicographic work in collaborative environments. The system also comes with a synchonization service which allows users to share and collaborate on document files, folders, and databases with other counterparts regardless of physical location. For the time being, we are developing only the web application version, but in the future, it is possible to also expand it into desktop and mobile applications.

Keywords: collaborative workspace; corpus tool; interactive lexicon viewer and editor, lexicographic application; Indonesian languages

1. Introduction

In modern lexicography, the use of computers in each step of dictionary-making process is inevitable. According to Atkins and Rundell (2008: 112), not until the beginning of 1990's did lexicographers begin to work directly on computers. Currently, it is very common to use electronic corpora in the development of lexicographic works, for example, in dictionary making. Many leading publishers have thus taken advantage of electronic corpora. The number of tokens in corpora has also increased over the years. If the corpus which was compiled for the COBUILD project consisted of eight million words, the current corpus, namely Bank of English, for the COBUILD project contains 4.5 billion words.

The use of information technology in lexicographic work is not only limited to the use of electronic corpora in dictionary making, but also extended to the use of computer tools in compiling and writing a dictionary. Atkins and Rundell (2012) mention that there are two types of software in lexicographic work. First, a corpus query system that enables us to analyze the data in a corpus in various ways and second is a software called dictionary-writing system (DWS) that enables lexicographers to compile and edit dictionary text. These lexicographic applications include all-key factors in making a dictionary, such as data collection, data analysis, and synthesis/composition. The advantage of lexicographic applications is having lexicographers to focus more on their expertise in compiling and writing a dictionary. Not only that, it has also cut and saved a lot of time and effort in the process of dictionary making.

The presence of DWSs has also given a new hope to under-resourced languages. According to Prinsloo (2012), under-resourced languages generally experience a lack of high standard dictionaries. Actually, less-resourced languages, such as many languages in Indonesia, also face deficiencies in language description and codification, including standard grammar, spelling guidelines, etc. Although in Indonesia there is a national agency for language affairs which oversees language development and conservation at a national scale, there still remains a number of languages which are under-resourced and less-described. Indonesia is the second-most lingustically diverse country in the world, with 719 languages spoken in the country. Among these languages, 386 languages have 5000 speakers or fewer. Most of them are now facing various degrees of language endangerment (Ethnologue 2015).

This paper concerns our attempts to develop a web application providing lexicographers in Indonesia with a shared workspace. This workspace is an inter-connected environment in which all the participants in dispersed locations can work and collaborate with each other in a single entity. Our DWS, called Lexcoworks, is mainly a web-based application. It means that the Lexcoworks interface can be opened with any regular browser. Furthermore, the Lexcoworks network feature makes it well-fitted for collaborative lexicographic projects, whether in a Local Area Network or on the Internet. In the case of using the Internet, people can access the project and do a lexicographic project from anywhere. It is also important to mention here that Lexcoworks is a multiuser DWS. It means that different users can log in simultaneously and work on the same project.

With regard to the nature of Lexcoworks as a web-based application, it will be a huge advantage if most users were familiar with the Internet. In Indonesia, the number of the Internet users still exhibits significant growth. However, there still remains two matters with regard to the number of Internet users in Indonesia. According to APJII, most Internet users in Indonesia regularly access the Internet through their mobile phones. Second, in Indonesia, Internet network coverage areas are still lacking compared to neighboring countries like Singapore and Malaysia. As a result, there remains many areas of Indonesia where Internet connection is not available. As a result, to anticipate various conditions of the Internet connection in Indonesia, Lexcoworks is equipped with a file synchronization feature to ensure that computer files in two or more locations are updated in certain rules. In this way, users can work on their project regardless of the availability of an Internet connection.

As a geographically-divided country, Indonesia has a primary geographic challenge related to the distance between its myriad islands. We develop our own system from scratch to meet the basic needs of lexicographers in Indonesia who are scattered in many local cities but prefer working in a team. To help people from remote and scattered areas to become involved and engaged in a common lexicographic project, Lexcoworks is developed as a shared workspace application.

A dictionary writing system which is designed and developed to meet our own demands is more applicable and adjustable than any of the existing systems. For this reason, we decided to develop our own system. The presence of Lexcoworks will be a great help in accelerating the process of language documentation and codification. Also, this shared workspace will help Indonesian lexicographers to conduct collaborative works and to solve their geographic obstacles.

2. Dictionary Writing Systems: An Overview

In this section, we review several existing DWS software programs. In the 1990's, the big dictionary publishers in the UK had already implemented DWSs in their dictionary projects with the aim of making dictionary compiling easier. According to Atkins and Rundell (2008: 112–114), a DWS comes with a ranging version from a simple and more elaborated program. A commercial DWS program is developed to meet with the dictionary publisher's qualification and demand. This kind of software should have the ability to manage the entire process of producing a dictionary, from compiling the first entry to outputting the final product for publication in printed or electronic media. Aside from DWS, this kind of software is also referred by other terms including 'dictionary editing system' (Svensén, 2009: 422), 'dictionary compilation software', 'lexicography software', 'dictionary production software', (De Schryver & Joffe, 2006: 41; Joffe and De Schryver, 2004: 17), 'lexicographic workbench' (Ridings, 2003: 204), 'dictionary management system' or 'lexicographer's workbench' (Langemets et al., 2010: 425), 'dictionary editing tool' (Krek 2010: 928), or 'dictionary building software' (Mangeot 2006: 185).

In general, Abel (2012: 87–88) distinguishes three main characteristics of a dictionary writing system. First is the content of the dictionary. Second is related to the structure or the grammar of the dictionary. The third aspect is the data presentation which includes formatting and style (see also De Schryver & Joffee, 2006: 41). Abel suggests these three aspects to be considered individually, but specific programs are best suited to work on each of them. Although Abel considers a DWS as an independent item, it could also be regarded as a system that takes benefit from other applications.

Basically, dictionary writing comprises mainly entry-inputting and editing that can happen in many different ways. This work can also be processed by using available word-processing systems that allow dictionary text to be processed and stored linearly, in exactly the way as it should be presented in the final product (Abel, 2012: 88). In addition, to separate the works of data-entry and data-editing, a lexical database can also be implemented, where the data are structured and stored in records, as well as separated from the emerging dictionary text. According to Svensén (2009: 421), from the point of view of the dictionary producer, such a database has the advantage of generating a great variety of products based on one and the same material. For this reason, our DWS is designed to be implemented with a database management system.

Furthermore, Abel (2012: 88) also mentioned that the use of mark-up languages also offers a significant help in dictionary writing. Mark-up languages, such as the popular XML, and editing software for them allow lexicographers to manipulate and manage documents in a structured way by adding additional information to the text in the form of tags; that is, standardized labels. Such kinds of mark-up languages are very helpful in lexicographic projects, but Abel wrote that we still need an additional tool to take benefit of the tags. Although many efficient and popular programs are available, these generic tools do not necessarily meet the needs of complex dictionary projects, because they were not specifically designed for lexicographic work (Abel, 2012: 88–89). We still need more specific tools: either it is an in-house tailor-made or off-the-shelf applications because dictionary projects are complex.

Atkins and Rundell (2008: 114) mention that a typical DWS consists of three main components: a text-editing interface, a database, and set of administrative tools. With a text-editing interface, lexicographers are able to create and edit dictionary texts. A dictionary database is required to store all the emerging dictionary text. Meanwhile, administrative tools support lexicographers to manage the project and publication process. According to Abel (2012: 95) a DWS is sometimes a written in-house system, such as an XML-editor customized for one or more dictionary project, or, in other cases, an off-the-shelf dictionary writing system package.

Not surprisingly, most software in Table 1 offer three components mentioned by Atkins and Rundell (2008), that is, a text-editing interface, a database, and administrative tool packages.

SIL has produced and launched their DWSs, namely, FLEx and Lexique Pro. These two software packages are robust lexical management systems that are suited to use in fieldwork and language documentation. In addition, Lexique Pro has a variety of tools from data entry and publication. EELex is an application that is built to manage Estonian languages. Among these DWSs, only four are entirely web-based, that is, *DEB2*, *Glossword*, *Lexonomy*, and *Mātāpuna*.

Dictionary Writing System	Description
EELex	a DWS developed at The Institute for Estonian Language (Eesti Keele Instituut) (Langemets et al. 2010)
FLEx	This software is produced by SIL International (SIL) for organizing and analyzing linguistic and cultural data. It enables linguists to be highly productive when building a lexicon and interlinearizinag texts.
TschwaneLex	TLex (aka *TshwaneLex*) is a professional, feature-rich, fully internationalised, off-the-shelf software application suite for compiling dictionaries or terminology lists.
Glossword	The software is aimed at creating online multilingual dictionaries, glossaries, references. It can mix several languages in one definition and create dictionaries written in different languages, managed by a single Glossword installation.
Lexique Pro	The software is an interactive lexicon viewer and editor, with hyperlinks between entries, category views, dictionary reversal, search, and export tools.
Lexonomy	A web-based DWS developed by Michal Boleslav Měchura which has the right balance between power and ease of use. It is designed to be a tool for writing and publishing dictionaries (and other dictionary-like datasets) where users find the right balance between power (= empowering users to do what they need to do) and ease of use (= not having a steep learning curve).
Mātāpuna	It is an open-source web-based DWS developed by Dave Moskovitz of Thinktank Consulting Limited in collaboration with the Māori Language Commission of New Zealand. (Bah 2010).

Table 1: List of selected Dictionary Writing Systems

Glossword is an open source tool written in PHP and intended for creation and publishing of an online multilingual dictionary, glossary, or reference. It means that Glossword only focuses on online dictionary writing. DEB2 is a web-based application with several features such as the server running in Linux, but clients are multiplatform. However, DEB2 appearance is still so basic that users need to improve their computer skills to deal with the applications. Lexonomy is a new DWS and still an experimental prototype, which will have some (probably not all) of the new and/or improved features described in the first three sections of this document (entry editing, dictionary configuration, publishing). Finally, Mātāpuna is developed as a web-based application that offers not only entry editing, dictionary configuration, and publishing, but also shared workspaces for collaborative work.

In terms of the availabilty of DWS basic components, our proposed DWS is quite similar to Mātāpuna. Like Mātāpuna, our system is also entirely web-based and does have functionality for collaborative work. However, due to the Internet network fluctuation, our system also implements file synchronization to anticipate it. This feature will prevent lexicographers from feeling frustrated when the Internet connection is lost, as is still the case in developing countries like Indonesia. In our system, we use character encoding UTF-8 that is capable of encoding all possible characters to accommodate any dictionary project in a different language. In the future, we are planning to extend the functionalities of our DWS by integrating a corpus query manager into our system.

3. Architecture and Functionality

In this section, we explain Lexcoworks architecture and functionality.

3.1 Design and architecture

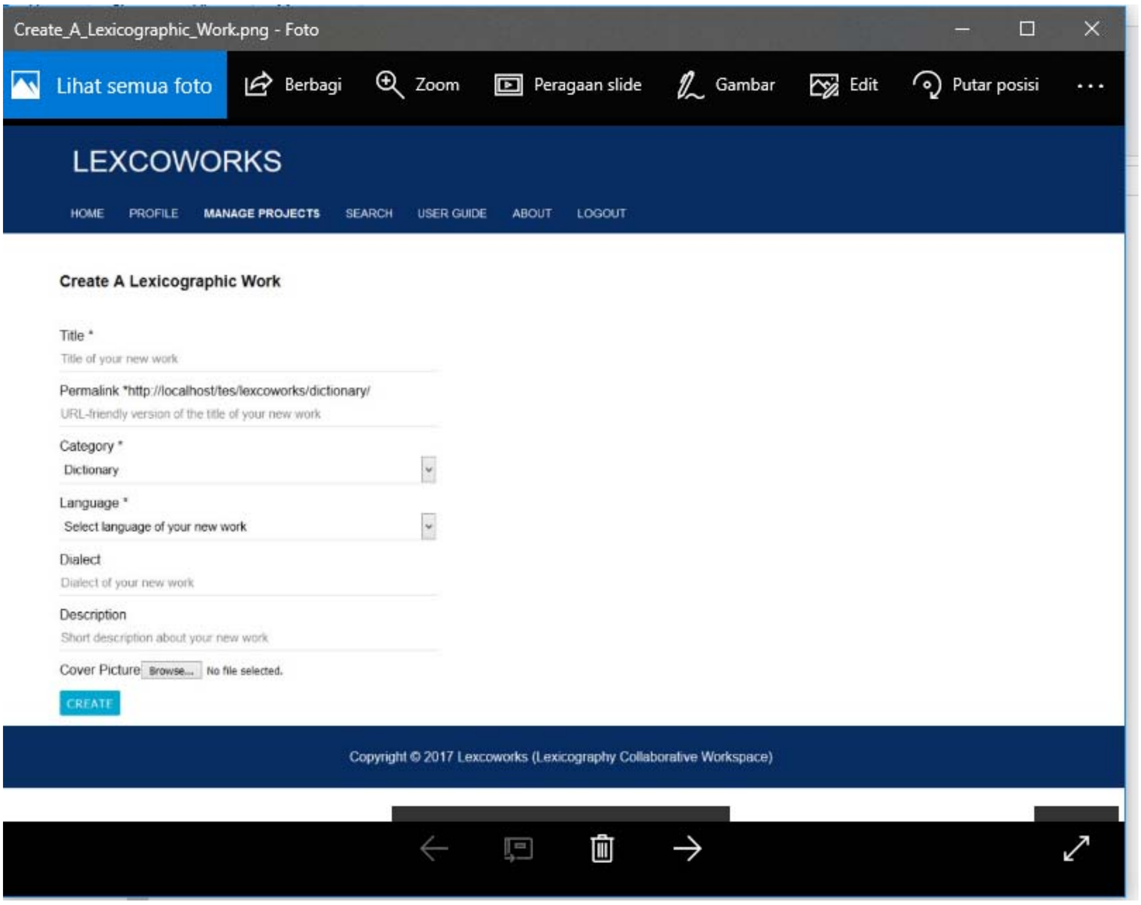
With regard to users, this application design is divided into two main types: administrator and non-administrator (uncategorized user). Uncategorized user refers to a user who does not yet have any role in a lexicographic project.

In addition to the two main types of user, there are three additional users including chief editor, editor, and contributor. These three user types represent user roles in lexicographic works including dictionary, thesaurus, and glossary. An uncategorized user can be either a chief editor, editor, or contributor in more than one lexicographic project. Users can have different roles in different lexicographic projects. For instance, a given user can be an editor in a Javanese dictionary project and at the same time s/he plays the role of contributor in a Madurese dictionary project.

An uncategorized user automatically becomes a chief editor whenever s/he starts a lexicographic project. Once an uncategorized user has created a lexicographic project through the create feature, s/he has the authority to use all features related to the project. Meanwhile, an uncategorized user can also play a role of an editor or contributor in a given language dictionary project, when a chief editor assigns him/her to the project. An editor is considered to be part of the core members of the team in a lexicographic project, so that s/he has also an authority to use features for dictionary building, but in a more limited way. On the contrary, although a contributor is also assigned by a chief editor to a lexicographic project, s/he does not belong to the core members of the team. For this reason, a contributor can only access and use very limited features that are related to a lexicographic project. An uncategorized user can be invited and promoted by a chief editor to become a contributor. At the same time, users can also submit an application to join in a dictionary project as contributors.

3.2 Functionality

Most features of Lexcoworks are available to users who are logged in. Users who are not logged in can use the search feature and view lexicographic works that have been published online on Lexcoworks. Users can sign up to Lexcoworks to get an account. Once a user signs up, s/he can log in to Lexcoworks as an uncategorized user. Users who are logged in can start or create a lexicographic project through the Create feature, join an existing lexicographic project, and manage their lexicographic projects in addition to searching and viewing lexicographic work features. Users who are logged in to Lexcoworks have more options available to them in the search feature, and can view lexicographic works that are not published online yet, long as they have the role of a chief editor, editor, or contributor of the lexicographic work.



Create A Lexicographic Work

Title *
Title of your new work

Permalink *http://localhost/tes/lexcoworks/dictionary/
URL-friendly version of the title of your new work

Category *
Dictionary

Language *
Select language of your new work

Dialect
Dialect of your new work

Description
Short description about your new work

Cover Picture Browse... No file selected.

CREATE

Copyright © 2017 Lexcoworks (Lexicography Collaborative Workspace)

Figure 1: Screenshot of Create Menu in LEXCOWORKS

The Create feature allows users to start or create a lexicographic project. Users can provide basic information about the project including title, category (whether it is a dictionary, thesaurus, or glossary), language, and cover picture. Once a lexicographic project is created, Lexcoworks will provide users with two pages consisting of an online page and an editing page of the project. The online page displays basic information about the project, lexicographic entries, a list of the team members of the project, a link to access the editing page (only available to the chief editor, editors, and contributors of the project), and a feature that allows users to submit a request to join the project as one of the contributors (available to uncategorized users who are logged in to Lexcoworks). If a lexicographic work has been published online, everyone can search and view its online page. If a lexicographic work has not been published online, however, only the members of the project can search and view its online page.

The editing page is a workspace for specific users to compile, edit, and add information with regard to the lexicographic project. A user can access editing features on the editing page. The availability of a feature depends on the user's role in the project. For instance, the feature to delete the project is only available to the chief editor; features to confirm entry change suggestions, publish a lexicographic work online, and print a lexicographic work are available only to the chief editor and editor; and features to edit an entry are available to all team members of the project. Lexcoworks allows users to add and edit entries online and offline. Users can add and edit entries online by filling and submitting a form directly on the editing page. For the web application version of Lexcoworks, if a user wants to add and edit entry offline or it is not possible to add and edit entry online (such as due to bad Internet connection or limited electricity access), a user can dowload a template file (tsv format) for offline editing. Users can add and edit an entry in a template file without Internet connection. Once users are connected to the Internet, they can upload the file to Lexcoworks. In addition to editing features, users can also access features to view the activity log on the editing page. The chief editor and editor can view the log of all activities in the project, such as which user edited that entry at what time, while a contributor can view only a log of his/her own activities.

Lexicographic entries of a lexicographic work in Lexcoworks could be divided into three different entities, that is entry, lemma, and sense. The entry entity represents a lexicographic entry. A lexicographic entry is assumed to consist of lemma (head word) and sense (meaning). Therefore, the entry entity includes information about identity number of the lemma entity, and identity number of the sense entity. The entry entity also includes information about the entry status, i.e. whether it is published online or not. If a lexicographic entry of a lexicographic work is published online, it is displayed on the online page of the work and can be searched by everyone. A lexicographic entry that is not published online can only be viewed on the editing page and can only be searched by members of the project.

The lemma entity represents a headword. It includes information about lemma name, lemma name with hyphenation point, pronunciation, word type, morphological structure, and homonym number. Information about word type is to indicate whether a lemma is a base or derivative. If a lemma is derivative, lemma entity also includes information about the identity number of its base. Information about morphological structure is to indicate morphological process, such as reduplication or affixation. Users can choose a morphological structure from a default set of morphological structures provided by Lexcoworks, or suggest a new morphological structure that is not included in the default set.

The sense entity represents a unit of meaning. It includes information about part of speech (such as verb, noun, and adjective), register (such as slang and formal), field (such as Chemistry and Biology), definition, example, and polysemy number. Users can choose from a default set of part of speech provided by Lexcoworks, or can suggest a new part of speech that is not included in the default set yet. Users can do the same for register and field labels; that is, choosing from default sets or suggesting a new register or field label. A sense entity is connected to a lemma entity by an entry entity. A lemma entity can be connected to more than one sense entity.

An entry entity belongs to an entity that represents a lexicographic work, namely opus entity. Therefore, the entry entity also includes information about identity number of the opus entity. The opus entity includes information about title, category (to indicate whether it is a dictionary, thesaurus, or glossary), language, cover picture, short description about the lexicographic work, identity number of the user who created the lexicographic work, and the time when the lexicographic work was created. Users can choose a language from a default set of languages provided by Lexcoworks or can suggest a new language. Languages that are included in the default set provided by Lexcoworks are referring to a list of languages from SIL.

For the time being, we have focussed on developing features for dictionary creation; therefore, the option to create a thesaurus and glossary is not available yet, and we only developed the web application version. The web application version of Lexcoworks was developed using custom PHP framework that we created, JavaScript for some functions, CSS for web design, and MySQL for database. We created custom PHP framework for Lexcoworks using model-view-controller (MVC) architectural pattern and UTF-8 character encoding. The web application version has been developed on localhost using XAMPP for Windows. The web application version will be hosted on cPanel shared web hosting. In the future, it is possible to expand Lexcoworks into a desktop application version and mobile application version.

4. Conclusion

The use of DWS is inevitable in lexicographic works. DWSs have become applications that include a range of components and modules with a great number of functions to deal with the complexity of dictionary making. Most DWSs offer three components including data-entry and editing, a database, and a set of administrative tools for publication. Our system, Lexcoworks, has all of these components including data entry interface, lexical database, and administrative tools.

We developed Lexcoworks rather than using existing systems to meet the basic needs of Indonesian lexicographers who wanted to work collaboratively from scattered remote areas. For this reason, we developed Lexcoworks as a web-based application, so it can be accessed through any web browser. The mission of Lexcoworks is to provide lexicographers with a shared collaborative workspace. The system allows users to work online and offline. If users want to add and edit entries offline or it is not possible to add and edit entries online (e.g. due to bad Internet connection or limited electricity access), users can dowload a template file (tsv format) for offline editing. Users can add and edit entries in a template file without Internet connection. Lexcoworks is designed to support multilingual dictionary projects, thus it uses character encoding UTF-8 to accomodate all possible characters.

With all these features, Lexcoworks can be a great help for lexicographers in multilingual Indonesia. Its users can work on lexicographic tasks anytime and anywhere, online as well as offline. They can do collaborate work in a more friendly and convenient environment because our system is regularly adjusted and revised to meet the Indonesian lexicographers' needs. In the future, we are planning to extend the functionalities of our DWS by integrating a corpus query manager into our system.

5. References

Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Bah, Oumar. (2010). Matapuna Dictionary Writing System: from Thinktank Consulting Limited. *Language Documentation & Conservation*, Vol. 4 (2010), pp. 169-176. http://nflrc.hawaii.edu/ldc/, http://hdl.handle.net/10125/4477

Kilgariff, A. (2006). Word from the Chair: In G.-M. De Schryver (ed.). DWS 2006: *Proceeding of the Fourth Internasional Workshop on Dictionary Writing System 7.* Pretoria: (SF)^2 Press.

De Schryver, G. M., & Joffe, D. (2006). The users and uses of TshwaneLex One. In *4th International Workshop on Dictionary Writing Systems* (DWS-2006) (SF) 2 Press, pp. 41-46.

Abel, A. (2012). Dictionary writing systems and beyond. In S. Granger & M. Paquot (eds.). *Electronic Lexicography.* Oxford: Oxford University Press, pp. 83-106.

Svensén, B. (2009). A Handbook of Lexicography: The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press.

Langemets, M., Loopman, A., & Viks, U. (2010). Dictionary management system for bilingual dictionaries. . In S. Granger & M. Paquot (eds.) *eLexicography in the 21st Century: New Challenges, New Applications.* Louvain-la-Neuve: Presses universitaires de Louvain.