Your language is easily recognized: A language clustering software in linguistic analysis

L R Nurjamin¹*, Y N Fajriah¹, A Hamzah¹, P Setiawati¹ and L Novita²

¹ Prodi Pendidikan Bahasa Inggris, Institut Pendidikan Indonesia, Jl. Pahlawan No. 32, Garut 44151, Jawa Barat, Indonesia
² Departemen Pendidikan Kesejahteraan Keluarga, Universitas Pendidikan Indonesia, Jl. Dr. Setiabudi No. 229, Bandung 40154, Jawa Barat, Indonesia

*luckyrahayu@institutpedidikan.ac.id

Abstract. The development of technology in upholding any discipline of knowledge has been widely extended, particularly in linguistic study. One of the advancement is a language clustering software, functioning to analyse any linguistic data through the sorting procedure, namely Monoconc. Due to its significance, this study is aimed at expounding the practicality and feasibility of this software. To achieve this aim, both qualitative and quantitative research approaches were employed. As results, Monoconc software is recommended to utilize in analysing linguistic corpora. The simplicity, as well as the viability, are the main roots of proposing this concordance software as an alternative to applying.

1. Introduction

The development of technology has been unfolded to any knowledge discipline, peculiarly in linguistics study [1]. Specifically, corpus linguistics is often associated with the use of the computer as the analysis tool of the corpus [2]. Lined up with this, Kirk highlights the role of the computer itself; that is, not merely as the data storage tool, but also as the device which can be operated to do editing, estimating, transforming, as well as showing the data in a certain format [3].

Moreover, an offered software to proceed the wording system is called as Monoconc. This application is a word clustering software which functions to get the key essence of the writers’ work. This corpus tool is often chosen by corpus linguists since it has a feature to figure out what they are interested in analyzing the corpus; that is, word frequency [4]. Furthermore, in his review, Barlow explains that concordance Monoconc program is considered effortlessly followed and navigated by the users since they do not need to find the particular option or setting in loading the data or searching some terms [5]. More than that, this concordance platform is suitable for a teacher and students when they are going to discuss particular issue coming up from the corpus by using the split screen in expanding the context [6].

Numerous related studies have been conducted. One of them was done by Tu and Wang whose study focused on analyzing tenses and rhetorical structure in journal article abstracts by using Monoconc and Wordsmith [7]. The result reveals that reporting verbs are mostly used in the texts, while the rhetorical moves used involve CARS model, IMRD structure, and IPMPRC structure. Another study was also undertaken in EFL teaching context, specifically, this concordance software is used to teach vocabulary...
to students [8]. It is shown that teaching vocabulary by benefiting the language corpora in this software is an effective and attractive way to do.

From the explanations above, it is clear to say that the studies showing the feasibility as well as the practicability of Monoconc software are very few. Hence, this study is an attempt to investigate how Monoconc platform is feasible and practically used in corpus linguistic studies.

2. Methods
This is categorized as a qualitative research since this study intends to describe a phenomenon and its characteristics; how Monoconc software is practically operated [9]. Additionally, this study was completed through some stages; preparing MonoConc software, collecting linguistics corpora, analyzing data (corpus), interpreting the usability of Monoconc, and drawing conclusions. As an extension, in this corpus software, the data processing employs descriptive statistics in which some calculation techniques are offered; frequencies, percentages/proportion, scale and others. To give clear description on how this corpus analysis tool works, six journal articles are used as the samples whose topic are related to a particular political issue in Indonesia. Monoconc software itself was downloaded from https://www.monoconc.com [10].

3. Results and discussion
The result reveals that MonoConc application is used through some stages; they are starting MonoConc, loading and unloading corpus. In the process of uploading the text, the data must be in txt file which can be run by notepad.

Specifically, this software also provides some sections; concordance, word frequency, and collocation. Concordance provides the word lists accompanying the frequent word. For instance, the concordance from one article shows that most appearance word accompanying is the word ‘was’. This shows the pattern of the author writing which tend to present the token condition in the past time.

Figure 1. The first page of MonoConc.
Another sample article can be seen from the collocation following frequency number of words. This frequency can be displayed in two different types; frequency order and alphabetical order. The most frequent word in the article entitled ‘the representation of newspaper toward Indonesian political issue’ is the name of politician ‘Prabowo’ with 26 appearances. This appearance illustrates the author tendency on presenting the figure of politician toward the topic of his research.

Table 1. Frequency and lexical list of the first article.

| Count | Frequency in percentage | Word          |
|-------|-------------------------|---------------|
| 72    | 1.5776%                 | Prabowo       |
| 57    | 1.2489%                 | Jokowi        |
| 54    | 1.1832%                 | Itu           |
| 54    | 1.1832%                 | Ma’ruf         |
| 47    | 1.0189%                 | Presiden      |
| 41    | 0.8983%                 | Dalam         |
| 41    | 0.8983%                 | Persen        |
| 39    | 0.8545%                 | Dengan        |
| 38    | 0.8326%                 | Dari          |
| 37    | 0.8107%                 | Sandiaga      |

From the point of collocation, the advance collocation feature provides the words collocated with the frequent word. In this part, the number of words collocated can be set manually. This means a number of left or right collocation of the frequent word can be customized based on the need of researchers.
Figure 3. Customizing number of words collocated.

For instance, as its collocation, the token ‘Prabowo’ was collocated with 12 negative and 14 positive words. Thus, the writing tends to bring the positive tendency toward the frequent word.

The explanations above show that MonoConc is easily used by many users, involving researchers, teachers, and students. Consonant with this, Barlow states that this software provides different configurations; therefore, it is friendly to the novice users as well as easily adapted to those who need more advanced and complicated options [5]. This is also confirmed by the study conducted by Jezo that the provided authentic languages and the reliability of this program are very helpful for EFL students in learning vocabulary [8].

4. Conclusion
As a result, MonoConc software is recommended to use since its simplicity and its friendly interfaces. This tool also offers assortment of functions in both statistical methods and a range of language options. In addition, seen from the result of the corpora analysis, MonoConc supports linguists to recognize the tendency of people’s thought through the texts. Therefore, this software is a welcoming and an appropriate alternative for corpus linguists, especially for the novice ones, in contributing to advancing the linguistic research approach.

Acknowledgments
We acknowledged Institut Pendidikan Indonesia.

References
[1] Baker P 2006 Using Corpora in Discourse Analysis (London: Continuum)
[2] Jensen K E 2014 Linguistics and Digital Humanities: (computational) corpus linguistics *Mediekultur* 57 115-134
[3] Kirk J M 1996 Review of K. Aijmer & B. Altenberg (Eds.) (1991). English Corpus Linguistics: Studies in Honour of Jan Svartvik. London: Longman *Journal of English Linguistics* 24 250-258
[4] Anthony L 2016 A critical look at software tools in corpus linguistics *Linguistic Research* 30 2 141-161
[5] Barlow M 1999 Corpus software assessment *International Journal of Corpus Linguistics* 4 1 173-184
[6] Svedkauskaite A 2004 Software review *Style* 38 1 127-133
[7] Tu P N and Wang S P 2013 Corpus-based research on tense analysis and rhetorical structure in journal article abstracts *PACLIC* 27 102-107
[8] Jezo E D 2013 Using language corpus in teaching foreign language vocabulary *International
Multidisciplinary e – Journal 2 1 11-25

[9] Nassaji H 2015 Qualitative and descriptive research: data type versus data analysis Language Teaching Research 9 2 129-132

[10] https://www.monoconc.com