# TESTING OF SUPPORT TOOLS FOR PLAGIARISM DETECTION FOR THE JAPANESE LANGUAGE: TESTOP-J PROJECT

Tolga Özşen[1], Salim Razı[1], Özgür Çelik[1], Senem Çente Akkan[1], İrem Saka[1], Tatsuya Sakaue[2]

[1]*Çanakkale Onsekiz Mart University, Turkey*
[2]*Hiroshima University, Japan*

## Keywords

Plagiarism detection, text-matching software, software testing, Japanese language, ideographic languages

## Abstract

The acquisition and/or learning process of Asian ideographic languages such as Japanese, Chinese, and Korean as a foreign/second language (L2) has several complex layers which are not limited to linguistic or grammatical features. Dominant cultural dynamics severely shape such languages with unorthodox writing systems for students outside the *Kanji region*. As to Japanese, the acquisition process has more layers not only because it has three unique ideogram-based writing systems (*Hiragana, Katakana, Kanji*), but also has differences in writing procedures (e.g. orthographic rules, punctuation marks, numbering, mixed/combined wording, etc.). Moreover, the interaction in daily life with Japanese language and culture is extremely limited, particularly for the Japanese L2 learners who are outside of the *Kanji cultural zone*. Spending a lot of effort on understanding the language and its culture may leave very little energy to focus on the academic integrity framework, resulting in academic misconduct cases, either intentionally or unintentionally.

On the other hand, academic misconduct issues (detection techniques, systems, tools, prevention methods, and etc.) for the Japanese language have been addressed in a small number of studies in the last two decades. The intersection points of most of those are the population and material they focus on. The majority of those studies focus on university students` writing assignments aiming to identify similarities based on words (syllables/characters) (Fukaya et al., 2003; Odaka et al.) or sentences (Suzuki et al., 2009) to reach the plagiarized web source from the paraphrased texts (Takahashi et al., 2007), and to develop detection systems (Ueta & Tominaga, 2010). Besides all those, Sakai and Tsuruhara argue academic misconduct behaviors related to professionals (e.g. duplicate submission for conferences) and the positioning of plagiarism in Japan through sanctions of universities as well (Sakai & Tsuruhara, 2012). Apart from these studies, Weber-Wulff's (2010) emphasis on the importance of the encoding variables (i.e. JIS-Shift and UTF-8) in plagiarism detection for

particularly Japanese language, apart from linguistic variables, is an important criticism that should be taken into account.

In an ideographic language such as Japanese as L2, acquiring the basic level, particularly for non-Kanji region students, takes a relatively long time. Consequently, academic writing techniques and detailed information regarding the promotion of academic integrity at the undergraduate level can only be taught limitedly and superficially. JLT related academic integrity studies mostly focus on citing techniques and ethics have become more visible in recent years (Yamamoto, 2016; Yamamoto et al., 2014; Yamamoto & Nitsū, 2015).

As can be seen, the Japanese language both as L1 and L2 is still very untouched territory in terms of detecting and analyzing academic misconduct issues and educational/pedagogical aspects. This ongoing project called "TeSToP-J" is the Japanese language version of the original TeSToP (Testing of Support Tools for Plagiarism Detection) project (Foltýnek et al., 2020) aiming to simulate the actual usage of text-matching tools in an educational setting by using a large collection of documents prepared in the Japanese language.

In this study, the methodology and protocols used in the original TeSToP project are revised in accordance with the characteristics of the Japanese language. This project aims to analyze Japanese-written documents compiled from four different sources (Wikipedia, online & open access papers, non-online materials, and multisource [Wikipedia & government white papers & OA papers]) by comparing several regional (Japan-based) and international web-based similarity detection tools using two main criteria (coverage and usability) as in original TeSToP test.

In order to test those systems and/or tools, seven disguising techniques (*copy & paste*, *paraphrase*, *synonym replacement*, *same content with different writing system* [e.g. *Kanji* instead *Hiragana, Hiragana* instead *Kanji*], *translation*, *stylistics* [white characters, images, etc.] and *encoding application* [UTF-8 and JIS-Shift]) will be used. Each document will be available in PDF, DOC, and TXT form. Using the original TeSToP methodology will allow us to classify those systems into categories from *useful systems* to *unsuited systems for the Japanese language*.

Taking into consideration the number of systems to be tested, the variety and number of testing documents, this ongoing project has a serious potential to be the most inclusive test on the Japanese language ever done. With the possible results obtained from this study, it is expected to contribute to all stakeholders such as vendors, professionals (academics), and decision makers in educational institutions. More importantly, we hope that this work, with its results, will be a source of inspiration for other ideographic and/or Asian languages too.

## References


Foltýnek, T., Dlabolová, D., Anohina-Naumeca, A., Razı, S., Kravjar, J., Kamzola, L., Guerrero-Dib, J., Çelik, Ö., & Weber-Wulff, D. (2020). Testing of support tools for plagiarism detection. *International Journal of Educational Technology in Higher Education, 17*(46), 1-31. https://doi.org/10.1186/s41239-020-00192-4

Fukaya, R., Yamamura, T., Kudō, H., Matsumoto, T., Takeuchi, Y., & Ohnishi, N. (2003). Hindo tōkei to gainen jisho wo mochiita bunshō no ruijisei no teiryōka [Measuring similarity between documents using term frequency and concept dictionary]. *Jōhō shori gakkai kenkyū hōkoku [IPSJ SIG Notes]*, 153, 73-79.

Odaka, T., Murata T., Gao, J., Suwa, I., Kuroiwa, J., & Ogura, H. (2003). n-gram wo mochiita Gakusei repooto hyōkashuhōnoteisatsu [A proposal on student report scoring system using n-gram text analysis method]. *IEICE, J86-D-1*(9), 702-705.

Sakai, Y., & Tsuruhara, T. (2012). Ronbun tōkō nı kakawaru hyōsetsutō no mondai nitsuite kōsatsu [Consideration about problems, such as plagiarism concerning paper contribution]. *Fundamentals Review, 5*(3), 239-243.

Suzuki, K., Takahashi, I., Shirai, H., Kuroiwa, J., Odaka, T., & Ogura, H. (2009). Hyōsetsu repooto hakkenn ni riyō suru 1 buntan`I de no kensaku kuerisakuseishushō [Web search query in detecting plagiarism reports]. *IEICE, J92-D*(11), 20172-2076.

Takahashi, I., Miyakawa, K., Odaka, T.,Shirai, H., Kuroiwa, J., & Ogura, H. (2007). Web saito karano hyōsetsurepooto hakken shien shisutemu [A Computer Aided Detection System for Learners` Reports Plagiarism from Web-site]. *IEICE, J90-D*(10), 2989-2999.

Ueno, S., Takahashi, I., Kuroiwa, J. Shirai, H., Odaka, T., & Ogura, H. (2006). Fukusuu no web peeji kara hyōsetsu shita repooto no hakken shien shisutemu no jissō [Implementation of a support system to find out of the report plagiarized from several web pages]. *Jōhō shori gakkai kenkyū hōkoku [IPSJ SIG Notes], 87,* 41-46.

Ueta, K., & Tominaga, H. (2010). A development and application of similarity detection methods for plagiarism of online reports. *Proceedings of ITHET 2010* (pp. 363–371).

Weber-Wulff, D. (2010). *Plagiarism detection test 2010*. https://plagiat.htw-berlin.de/software-en/2010-2/

Yamamoto, F. (2016). Ronbun no 'itoteki dehanai hyōsetsu' no mondai: modaritii no kondō to kaishaku no nai in`yō [Unintentional plagiarism in Japanese writing: Confusion of modalities and citation without interpretation].*Global Communication, 6,* 117-132.

Yamamoto, F., & Nitsū, N. (2015). Ronbun no in`yō – kōzō: Jinbun&shakaikagakukei ronbun shidō no tame no kisoteki kenkyū [Quotation and interpretation structure of literature-analysis papers: Basic research on instruction for writing papers in humanities and social science]. *Nihongo Kyōiku [Journal of Japanese Language Teaching], 160,* 94-109.

Yamamoto, F., Nitsū, N., Ohshima, Y., & Satō S. (2014). In`yō kara kaishaku ni itaru in`yōbun no tayōsei [Varieties of citations from quoting to interpreting in the "literature-analysis-type" papers in humanity and social science]. *Dai 16 kai Senmon Nihongo Kyōiku Gakkai Ronshū [Conference proceeding of 16th conference of the society for technical Japanese education]* (pp. 16-1).