Al-Thwaib et al. International Journal of Educational Technology in Higher Education (2020) 17:1
https://doi.org/10.1186/s41239-019-0174-x

International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE Open Access

An academic Arabic corpus for plagiarism detection: design, construction and experimentation

Eman Al-Thwaib[1], Bassam H. Hammo[2*] and Sane Yagi[3]

* Correspondence: b.hammo@ju.edu.jo
[2]Computer Information Systems Department, King Abdullah II School of Information Technology, University of Jordan, Amman, Jordan
Full list of author information is available at the end of the article

Abstract

Advancement in information technology has resulted in massive textual material that is open to appropriation. Due to researchers' misconduct, a plethora of plagiarism detection (PD) systems have been developed. However, most PD systems on the market do not support the Arabic language. In this paper, we discuss the design and construction of an Arabic PD reference corpus that is dedicated to academic language. It consists of (2312) dissertations that were defended by postgraduate students at the University of Jordan (JU) between the years 2001–2016. This Academic Jordan University Plagiarism Detection corpus; henceforth, JUPlag, follows the Dewey decimal classification (DDC) in the way it is structured. The goal of the corpus is twofold: Firstly, it is a database for the detection of plagiarism in student assignments, reports, and dissertations. Secondly, the n-gram structure of the corpus provides a knowledgebase for linguistic analysis, language teaching, and the learning of plagiarism-free writing. The PD system is guided by JU Library's metadata for retrieval and discovery of plagiarism. To test JUPlag, we injected an unseen dissertation with multiple instances of plagiarism-simulated paragraphs and sentences. Experimentation with the system using different verbatim n-gram segments is indeed promising. Preliminary results encourage that permission be sought to enrich this corpus with all the theses in the Thesis Repository of the Union of Arab Universities. The JUPlag corpus is intended to function as an indispensable source for testing and evaluating plagiarism detection techniques. Since the University of Jordan is seeking to become a center for plagiarism detection for Arabic content and being a non-profit organization, it will charge a nominal fee for the use of JUPlag to finance the maintenance and development of the corpus.

Keywords: Corpus tools, Natural language processing, Plagiarism detection, Text plagiarism, Verbatim plagiarism

Introduction

Plagiarism is simply defined as appropriating others' words, thoughts, or intellectual property without providing proper citation or giving credit to them as the original source. The Oxford Dictionary[1] defines plagiarism as "The practice of taking someone else's work or ideas and passing them off as one's own". With the exceptionally large

[1]https://en.oxforddictionaries.com/definition/plagiarism

volume of articles, reports and books available on the Internet, plagiarism in academic writing is a major concern that has become the matter of the moment.

Plagiarism can be either intentional or unintentional (DeVoss & Rosati, 2002). It is intentional when copying or modifying someone else's words without providing proper citation to the original source. It is unintentional when one copies from others without knowing the rules and regulations for academic writing. However, ignorance should not be an excuse. For instance, the latest scandal of alleged plagiarism involved a respectable lecturer at an Ivy League university who once was the executive editor of a major newspaper. It cast doubt on the integrity and reputation of an otherwise highly respectable academic and public figure. This academic had properly credited alleged instances of plagiarism to their sources, sometimes repeatedly, but occasionally failed to do so. This 'unintentional plagiarism' is a form of academic dishonesty.

Advancement in technology both facilitates plagiarism and prevents it. At the click of a mouse, paper mill websites help students and researchers to copy or buy research papers. Yet, plagiarism detection systems deter the appropriation of others' intellectual property. Plenty of websites are nowadays offering tools for plagiarism detection. Some sites are commercial but few are free. Turnitin and PlagScan, for instances, are very popular commercial tools that are used world-wide for the detection of text plagiarism. They are capable of detecting different forms of plagiarism that range from simple copy-paste plagiarism to word switching, sentence and paragraph paraphrasing, etc. However, these tools do not prevent plagiarism but catch it after it has occurred (Beute, Van Aswegen, & Winberg, 2008).

Misconduct in Arabic research is not an exception. Unfortunately, however, most of the plagiarism detection tools act on ASCII (American Standard Code for Information Interchange) data and very few support Unicode data for plagiarism comparisons. Plagiarism detection for scholarly research written in the Arabic language is not well supported. The scarcity of Arabic literature and resources on the Internet as well as the shortage of commitment to research in Arabic NLP (Natural Language Processing) are the main reasons behind the absence of efficient plagiarism tools that support a language spoken and written by around 423 million people.

The main contribution of this ongoing project is twofold. At its preliminary stage, it will construct a plagiarism corpus made of defended dissertations in the thesis repository at the library of the University of Jordan. The second is to develop a plagiarism detection system dedicated to the Arabic language that is capable of detecting verbatim plagiarism and some intelligent plagiarism including word order changes, paraphrasing and synonym replacement. Hereafter, we refer to the corpus as JUPlag and to the plagiarism detection system as PD system.

The remaining of the paper is organized as follows. Section 2 provides a background and discusses related literature. Section 3 introduces the research methodology. Section 4 discusses the experiments and findings. Finally, section 5 presents the conclusion of this paper and future work.

Background and literature review

Plagiarism

The lack of fundamental research skills could be the common reason why university students/researchers plagiarize (Devlin & Gray, 2007). However, academic

writing is not an easy task. It requires clarity, conciseness, focus, structure, and evidence. It requires a lot of reading, appropriate usage of words and grammar, and learning how to express ideas and thoughts. Several studies pointed to other reasons for plagiarism: lack of author confidence, shortage of time, fear of failure, pressure of parents and scholarship committees to maintain high grades, lack of punishment by the institution, ease of appropriation, and absence of good plagiarism detection systems (Devlin & Gray, 2007; Eret & Ok, 2014; Franklin-Stokes & Newstead, 1995).

From a legal point of view, the act of plagiarism is not considered a crime (Frye, 2016). However, plagiarism during university years is highly condemned by the academic community and it may leave a significant impact on one's career beyond academia. "Consequences range from loss of reputation to economic fines and ruined careers. Students are expelled from their schools, and faculty fired... Doctoral degrees can be revoked and plagiarizing publications are retracted and cursed" (Satija & Martínez-Ávila, 2019, p. 90). A case in point is the disgrace of politicians (cf. Ruipérez & García-Cabrero, 2016).

Plagiarism is of seven types: paraphrasing a text without proper citation, mosaic plagiarism where text from different sources is combined into one, copy and paste without due citation, incorrect citation, arrogating someone else's entire work, self-plagiarism where one submits his/her published work as though it were new, and citing a non-existing work (Vij, Soni, & Makhdumi, 2009).

Plagiarism prevention methods have a long-term positive effect, but, unfortunately, their implementation is usually time-consuming (Lukashenko, Graudina, & Grundspenkis, 2007). Relying on such methods to maintain academic integrity, however, won't be enough to stop researchers from plagiarizing. In the words of Bolkan (2006), "Many educators blame the internet for what they perceive as the rise of plagiarism. Although the Internet certainly enables more efficient plagiarism, blaming it for widespread copying is akin to blaming a bank robbery on the presence of cash in the building … Efforts must be directed at prevention as well as detection and punishment. (p. 4)".

Plagiarism detection software (PDS) can be content-based (extrinsic) or stylometry-based (intrinsic) (Rahman, 2015). Extrinsic plagiarism detection (EPD) discovers instances of appropriation by comparing a suspicious document with reference documents (a database or a corpus). Intrinsic plagiarism detection (IPD), on the other hand, discovers instances of appropriation in the suspicious document without using any reference corpus. Figure 1 depicts the common types of text plagiarism and the classification of plagiarism detection software tools.

A plagiarism detection system has to ideally handle most types of plagiarism, including text modification by word-shifting, translation, and summarization that bypass string-matching tools. At this preliminary stage, our present work handles string-matching-based plagiarism detection and it is planned that it will be enhanced with such NLP techniques as stemming and part-of-speech tagging, and by the use of such lexical resources as the work of (Baras, Sawalha, and Yagi: A more extensive wordnet for Arabic, submitted), Arabic-WordNet,[2] dictionaries, and thesauri.

[2] http://globalwordnet.org/arabic-wordnet/

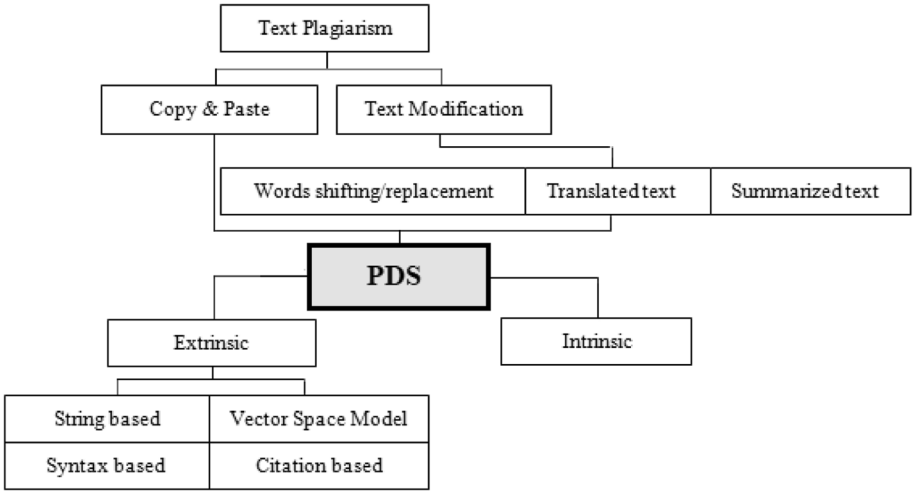

Fig. 1 Types of text plagiarism detection software tools

Related work

Plagiarism is an old topic and it has been well studied in the literature. In this section, we only focus on the recent work on Arabic text plagiarism detection. However, for further reading on the topic of plagiarism, we refer the reader to Maurer, Kappe, and Zaka (2006). In addition, the following is a sample of scholarly work that exemplifies plagiarism types with reference to Fig. 1. For intrinsic plagiarism, we refer the reader to the work of AlSallal, Iqbal, Palade, Amin, and Chang (2019), Polydouri, Siolas, and Stafylopatis (2017), Tschuggnall and Specht (2012), Zu Eissen and Stein (2006); for string-based extrinsic plagiarism detection, refer to Baba, Nakatoh, and Minami (2017), Leonardo and Hansun (2017), Nakatoh, Baba, Yamada, and Ikeda (2011), Wise (1996); for vector-space-based plagiarism detection, see Kong, Zhao, Lu, Qi, and Zhao (2016), Meuschke, Siebeck, Schubotz, and Gipp (2017), Paul and Jamal (2015); for syntax-based plagiarism detection refer to Si, Leong, and Lau (1997), Vani and Gupta (2017); and for citation-based detection see Gipp and Beel (2010), Gipp and Meuschke (2011); and Meuschke, Gipp, Breitinger, and Berkeley (2012).

The first shared task that addressed plagiarism detection in Arabic texts is "AraPlag-Det" (Arabic Plagiarism Detection) introduced in the PAN@Fire2015 competition and it has become since then an annual event that involved extrinsic and intrinsic plagiarism detection (Bensalem et al., 2015). Researchers in Arabic NLP adopted shared tasks to raise awareness of plagiarism problems and to develop solutions to them.

The majority of works on Arabic plagiarism detection involves preprocessing, segmenting documents into chunks of sentences of variable sizes (n-grams), tokenization, removing diacritics and non-alphanumeric characters, normalizing some letters (for example "ﺇ،ﺃ،ﺁ" get normalized into "ا"), stemming, lemmatization, part-of-speech tagging, and synonym replacement.

Zaher, Shehab, Elhoseny, and Osman (2017) developed a web-based plagiarism detection system for Arabic documents, called APDS. The system operated in three phases: preparation, preprocessing, and similarity detection. After preprocessing, the query document was presented as n-gram chunks for similarity detection. The proposed system was tested on a dataset of 10 Arabic documents and evaluated in terms of

precision and recall. The authors claimed an average precision of 82% and an average recall of (92.5%). However, the paper does not tell what kind of plagiarism was detected, how the documents were presented or how the precision and recall measures were obtained.

Mahmoud and Zrigui (2017) proposed a system for detecting semantic plagiarism in Arabic documents that benefited from machine learning technology. In the preprocessing phase, the suspicious and source documents were split into sentences then into words without removing stopwords. In the feature extraction phase, the TF*IDF (Term Frequency-Inverse Document Frequency) measure was calculated for weighting words in terms of importance. Then the *word2vec* algorithm was used for learning word embeddings, and the *skip-gram* model was employed for predicting the context of words given a current word vector. For similarity calculation, they used cosine and the Euclidean distance measures. The degrees of similarity between sentences were compared to a predefined threshold. Experiments were conducted on an open source Arabic corpus and they claimed a precision rate of (85%) and a recall rate of (84%).

Mahmoud, Zrigui, and Zrigui (2017) used a Convolutional Neural Network (CNN) approach for detecting paraphrasing plagiarism in Arabic documents. This method is said to detect paraphrasing plagiarism through the measurement of semantic relatedness between the suspicious and the original documents. Their approach has three phases: preprocessing, feature extraction, and paraphrase detection. After preprocessing, the feature extraction phase employed a *skip-gram* model for word-to-vector representation, where each document is represented by a vector in a multidimensional space. The paraphrase detection phase applied the cosine similarity measure on the vectors of both the suspicious and the original documents to reduce dimensionality. Finally, a mathematical function called Softmax was used for paraphrase detection according to some predefined threshold. Experiments showed a precision rate of (88%).

However, Mahmoud et al. (2017) and Mahmoud and Zrigui (2017) conducted their experimentation on an open source Arabic corpus, named OSAC (Saad & Ashour, 2010). The corpus was organized in ten different categories collected from multiple websites. The sources of the articles were news channels and social and commercial websites, which clearly makes it inappropriate for academic plagiarism detection. Specialized content is what the PD corpus ought to consist of, because academics do not normally plagiarize the news or social media.

Abdelrahman, Khalid, and Osman (2017) presented a framework for content-based PD in Arabic documents. Their framework has two phases: preprocessing and document representation. They used a tree-structure model with the document at the root of the tree, the paragraphs at the second level, and the sentences at the third level of the tree. A Longest Common Substring (LCS) matching algorithm was used for comparing hashed text chunks (i.e. words in their case). No experiments were made to evaluate the system or show its effectiveness and therefore there was no plagiarism detection corpus.

Ghanem, Arafeh, Rosso, and Sánchez-Vega (2018) presented a system for detecting extrinsic plagiarism in Arabic texts. Their system, Hybrid Plagiarism (HYPLAG), followed a hybrid detection approach. They adopted corpus-based and knowledge-based approaches for the detection of both the verbatim and rephrasing types of plagiarism. The system was compared to other systems that participated in the Arabic

Plagiarism Detection PAN-Forum for Information Retrieval Evaluation (AraPlagDet PAN@FIRE) competition and was tested on a corpus called External Arabic Plagiarism Detection (ExAraPlagDet-2015). The authors reported that HYPLAG outperformed others with a success rate of (89%). They chunked the query (suspicious) document and the source documents into n-term sentences. Then the synonyms of the query document were extracted from the Arabic-WordNet. The original sentences were ranked with respect to the suspicious sentences and the ones with the highest scores were extracted as potentially plagiarized sentences. Finally, the candidate sentences and suspicious sentences were compared for similarity using the vector space model and the TF*IDF weighting measure. A similarity value that exceeded a predefined maximum threshold indicated plagiarism, while a similarity value between minimum and maximum thresholds required a call for the next phase of feature-based semantic similarity measurement based on the synonyms extracted from the Arabic-WordNet.

Khorsi, Cherroun, and Schwab (2018) used a Two-Level Plagiarism Detection System (2 L-APD), which is said to detect different plagiarism cases, including verbatim and paraphrasing. Their system consisted of two consecutive modules: fingerprinting and word embedding detection. The first module is responsible for preprocessing and segmenting the suspicious document into sentences. When sentences exceeded some threshold value, they were passed on to the second module to test for paraphrasing and synonym replacement. The fingerprinting was applied by chunking the text documents into n-grams and then selecting the least frequent ones. Finally, they used a function called Brian Kernighan and Dennis Ritchie (BKDR) for hashing the selected n-grams. The first module applied Jaccard measuring similarity, whilst the second module used the cosine similarity measure. Important words were picked on the basis of their IDF value and their part of speech tags. To test their approach, Khorsi et al. (2018) used the ExAraDet-2015 corpus. Experimental results showed an overall precision rate of (85%) and a recall rate of (87%).

Although the works of Ghanem et al. (2018) and Khorsi et al. (2018) seem promising, they both have been tested on ExAraDet-2015 corpus, which is an Arabic corpus made of short sentences constructed for the PAN@FIRE plagiarism detection competition. We suspect this corpus might not be suitable for academic plagiarism detection as it is not a well-organized academic corpus, neither it is discourse-structure annotated.

Clearly, there is need for a corpus dedicated to plagiarism detection that is authentic, big, versatile, and richly annotated. The JUPlag corpus is intended to meet this need and to function as a test bed for the evaluation of plagiarism detection techniques.

Corpus design methodology

The JUPlag corpus was guided by the following design objectives:

1) To compile academic texts for the purpose of training and testing the Arabic plagiarism detection system that is to be developed.
2) To devise a mechanism for organizing the texts and indexing them.
3) To annotate the texts using a stemmer and a part-of-speech tagger.
4) To construct an Arabic thesaurus database that can be used for detecting synonym replacements.

Source data collection

Data collection is a fundamental success factor in plagiarism detection. PD systems need to access multitudes of sources of data to detect potential plagiarism. This includes accessing local databases as well as online data available on the internet. Due to the scarcity of scholarly Arabic literature that is in digitized form, it has been deemed necessary to build a resource that would contain a collection of academic texts, a resource that may be used for the detection of plagiarism in dissertations before a defense is scheduled. Postgraduate students usually sign an affidavit stating that they observed the code of ethics in the compilation of their theses, that they accepted all legal repercussions of plagiarism including the revocation of their degrees, and that they agreed that the Deans Council revocation decision would be final.

With the necessary legal provisions, the Library of the University of Jordan graciously gave us permission to access their copyrighted repository of dissertations. The University requires that postgraduate students transfer their copyrights to it and get them to sign an authorization form that permits the University of Jordan "to supply copies of [their] Thesis/Dissertation to libraries or establishments or individuals on request, according to the University of Jordan regulations". We have obtained permission of the University administration and of the Director of the University Library to access the dissertation repository for the specific purpose of the development of the JUPlag corpus and for experimentation with the repository.

We had access to (2312) dissertations that were defended by University of Jordan postgraduate students between the years 2001–2016. Table 1 shows the number of collected dissertations per year. Notice the significant increase in the number of collected dissertations in 2006 and beyond; this is due to the School of Graduate Studies' drive to boost the number of master's, doctoral and high specialization programs. As JU sought to become a pioneer in postgraduate programs, it widened its program offerings resulting in 2012 in (105) master's programs, (34) doctoral programs, and (16) high specialization programs in Medicine. As of today, the Graduate School offers (123) master's programs, (38) doctoral programs, (16) high specialization programs in Medicine, and (1) high specialization program in Dentistry.

Challenges identified

In the process of constructing the JUPlag corpus, the following problems were encountered:

Table 1 Per year distribution of the collected dissertations

Year	Count	Year	Count
2001	23	2009	203
2002	7	2010	261
2003	13	2011	174
2004	18	2012	141
2005	47	2013	187
2006	126	2014	317
2007	202	2015	282
2008	186	2016	125
Total = 2312			

1) **Differences in dissertation format and structure**

Although the school of graduate studies at JU has guidelines and a standardized template for dissertations, there are some variations among schools and disciplines. This might include the number of chapters, pages, dissertation layout, and fonts. For the past 10 years, a graduate student has been required by law to hand in an electronic copy of his/her dissertation upon its endorsement by the school of graduate studies. Prior to that, hard copies were submitted to the library whose staff had to retype the dissertations, a cumbersome and costly exercise.

Due to copyright law restrictions, we had to obtain permission to process the content of the repository for the purpose of constructing the JUPlag corpus.

2) **Scarcity of Arabic online literature**

The success of plagiarism detection is dependent mainly on access to online resources and on offline databases. Unfortunately, there is a limited volume of machine-readable Arabic scholarly articles online. Hence, testing our system will be restricted to JUPlag corpus. At a later stage, we will seek permission to include in this corpus all the dissertations in the repository of the Union of Arab Universities.

3) **Paucity of efficient Arabic tools**

Arabic suffers from the scarcity of free NLP tools. Tokenization, root extraction, part of speech tagging, and sentence boundary identification are essential for many NLP tasks. Root extraction reduces word tokens to word types. A Part-of-Speech Tagger (POST) is essential for machine translation, dependency parsing, and language pattern extraction. Online dictionaries, thesauri, and semantic networks are indispensable for meaning-centered tasks. Although many of these essential tools do exist, they are not available for free. Many of those that are free of charge are not reliable. Hence, researchers in the field of Arabic NLP often decide to build their own tools.

Construction of the Arabic academic plagiarism detection corpus

To the best of our knowledge, the only available extrinsic plagiarism corpus devoted to Arabic text plagiarism detection is ExAraDet-2015.[3] The corpus was used in the PAN@Fire2015 competition to judge and to rank the competing solutions. The corpus is made of 1171 short documents, of which (48.68%) are source documents and (51.32%) are suspicious. The following is a detailed description of our design and construction of JUPlag, the Arabic academic plagiarism detection corpus.

Corpus architecture

The architecture of JUPlag follows the Library of JU in the way it classifies its content. JU Library holdings are classified in accordance with the DDC system and it uses some standard metadata. The following is a brief description of the two classification techniques that we adopted while building the plagiarism corpus.

[3] http://misc-umc.org/AraPlagDet/?i=1

The Dewey decimal classification system

The DDC[4] system is the world's most widely used technique to organize library collections. It has been named after its founder, Melvil Dewey, an American Librarian who developed it in 1876. The DDC system represents an adaptive knowledgebase which is revised continuously to cope up with knowledge development. It has been developed and maintained by the Library of Congress. The DDC system has 10 main subject categories. Each category is represented by a three-figure value in the range from 000 to 999 (Chan, Comaroni, Mitchell, & Satija, 1996).

The JU Library had adopted DDC in the classification of its holdings, whether they are books, magazines, periodicals, or dissertations, etc. As Fister (2009) notes, "Dewey can sort large collections into more specific groups than BISAC can. (p. 24)".

A Dewey numerical scheme has three levels. Altogether, they make the classification number of a library item. Table 2 shows the first level categories. For instance, a dissertation about Arabic dictionaries "المعاجم العربية" would carry the Dewey number 413. The number can be interpreted as follows. Level-1 (400) is used for the *language* "اللغات" category, level-2 (10), a multiple of tens level, is used for the Arabic language "اللغة العربية" category, and the third level (3), a sequential number, is used for the Arabic Dictionary "المعاجم العربية" category.

JU library's metadata

In addition to using DDC for classifying its items, the JU Library also adopts a set of standard metadata for their classification. The metadata include: barcode, author's first name, author's surname, title, date of publication, subject, and the call number that specifies the shelf location of the item. Metadata are used to locate and retrieve information quickly.

An interesting characteristic of JUPlag is that its content is organized according to DDC system. This organizational structure is advantageous in that it categorizes theses/dissertations according to subject matter which makes it possible to perform plagiarism detection within a subcorpus rather than the entire corpus, a procedure that saves precious processing power and time. Search in one DDC category of theses/

Table 2 Dewey decimal classification system

Dewey #	English Categories	JU Arabic Categories
000–099	General References or Works	المعارف العامة (العموميات)
100–199	Philosophy, psychology & logic	الفلسفة وعلم النفس والمنطق
200–299	Religion	الديانات
300–399	Social Sciences	العلوم الاجتماعية
400–499	Language	اللغات
500–599	Natural Science	العلوم الطبيعية
600–699	Technology and Applied Science	العلوم التطبيقية (التكنولوجيا)
700–799	Fine Arts & Recreation	الفنون الجميلة
800–899	Literature	الأدب
900–999	History, Geography & Biography	التاريخ والجغرافيا والسير

[4]https://www.oclc.org/en/dewey/features/summaries.html

dissertations is also what linguistic analysis would do when they want to study the discourse characteristics of a genre or its embedded linguistic patterns.

In a similar manner, DDC has been successfully used by Jenkins, Jackson, Burden, and Wallis (1998) to automatically classify web resources and by Golub, Lykke, and Tudhope (2014) to enhance Information Retrieval (IR) and indexing systems.

Data processing outline of the JUPlag corpus

In this section, we describe the processing stages of the corpus construction. Figure 2 depicts the overall data processing stages. Table 3 shows the distribution of the corpus dissertations in accordance with the Dewey categories.

Tokenization

The tokenization process takes a dissertation *D* and splits it into separate words (unigrams). We designed and implemented a tokenizer that extracts words at multiple delimiters, including white spaces, tabs and punctuation marks (Hammo, Yagi, Ismail, & AbuShariah, 2016). The output of the tokenizer is of two types: tokens that correspond to units whose characters are recognizable such as punctuation marks, numeric data, dates, etc., and tokens that need further morphological analysis. Tokens of one or two-character length, non-Arabic characters, or numerical values are ignored and excluded from the database. Stop-words were also removed from the corpus. Developers of NLP applications usually remove stop-words from search engine indices as this will reduce

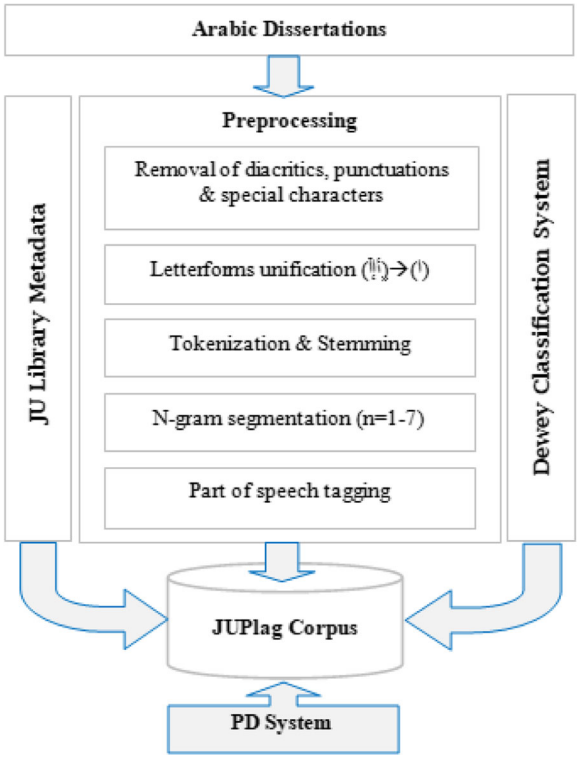

Fig. 2 Processing outline of the JUPLag Corpus

Table 3 Distribution of the corpus dissertations in accordance with Dewey's categories

Dewey #	English Categories	JU Arabic Categories	Number of Dissertations
000–099	General References or Works	المعارف العامة (العموميات)	17
100–199	Philosophy, psychology & logic	الفلسفة وعلم النفس و المنطق	10
200–299	Religion	الديانات	397
300–399	Social Sciences	العلوم الاجتماعية	1338
400–499	Language	اللغات	42
500–599	Natural Science	العلوم الطبيعية	17
600–699	Technology and Applied Science	العلوم التطبيقية (التكنولوجيا)	103
700–799	Fine Arts & Recreation	الفنون الجميلة والديكور	147
800–899	Literature	الأدب	134
900–999	History, Geography & Biography	التاريخ والجغرافيا والسير	107
Total			2312

the size of indices dramatically (Salton & Buckley, 1988; Yang, 1995) and that will improve recall and precision.

Segmenting dissertations into n-grams

For a given dissertation *D*, we split the sentences of *D* into n-gram segments. An n-gram segment is a substring of *n* consecutive words. The popular forms of n-grams include bi-gram (2 words), tri-gram (3 words), and four-gram (4 words). The maximum value we considered in preparing the corpus is $n = 7$ (seven-gram). The n-grams will be used later in a string matching algorithm to detect similarity between the source sentences and the suspicious ones.

Before the splitting process, punctuation, special characters, and diacritics get removed and letterforms normalized; i.e., all shapes of *alif* and *hamza* get converted to one form each. To explain how the n-gram segments were formed, consider the Arabic sentence "ذهب احمد الى السوق واشترى خبزا وعسلا" and its English translation, "Ahmad went to the market and bought bread and honey". A sliding window of size *n* splits this text as demonstrated in Table 4.

Stemming

Stemming is the process of mapping derivative words onto the base form, the stem, that they share. Stemming uses morphological heuristics to remove affixes from words

Table 4 N-gram segments generated from a sentence

Unigram	Bigram	Trigram	4-g	…
ذهب	ذهب احمد	ذهب احمد الى	ذهب احمد الى السوق	
احمد	احمد الى	احمد الى السوق	احمد الى السوق واشترى	
الى	الى السوق	الى السوق واشترى	الى السوق واشترى خبزا	
السوق	السوق واشترى	السوق واشترى خبزا	السوق واشترى خبزا وعسلا	
واشترى	واشترى خبزا	واشترى خبزا وعسلا		
خبزا	خبزا و عسلا			
وعسلا				

before indexing them. Arabic stemming is more complex than it is in English. Arabic is a morphologically introflexive, fusional language (Velupillai, 2012), whilst English is morphologically hybrid. Sapir and Swiggers (2008) label English as a mixed-relational fusional language. The majority of words in the Arabic language, on the other hand, are primarily constructed from three-consonant roots and a set of morphological patterns. With prefixes, infixes, and/or suffixes interdigitated with the root radicals, multitudes of words are derived. Then these coined words, if generated with verb patterns, get inflicted for number, gender, mood, voice, and tense; if generated with noun patterns, they get inflicted for number, definiteness, and case. An Arabic stemmer should identify the base word and remove all inflectional and derivational affixes. It should recognize, for example, that the strings, كاتب kAtib 'writer', كتاب kitAb 'book', مكتبة maktabatun 'library', as belonging to one root, كتب KTB 'to write'. For this task, we used Khoja and Garside's (1999) Arabic stemmer.

Part of speech tagging (POST)

A part-of-speech tagger (POST) is a software application that reads text in a particular language and assigns to each word its word category; i.e., it marks it as noun, verb, adjective, etc. Part of speech tagging is an essential process in understanding how sentences are formed from small constituents. It is mainly used in syntactic and semantic analysis of sentences. For this task, we used MADAMIRA,[5] a comprehensive tool for Morphological Analysis and Disambiguation of Arabic. Adding POS annotations to the corpus is mainly to prepare the corpus for the next stage of this ongoing project. Similar to the work of Elhadi and Al-Tobi (2008), we intend to use the part-of-speech tags to represent the structure of text segments for further comparisons and analysis. Plagiarized text tends to have the same POS tag features as the original source.

The final academic corpus

The final academic corpus (database) constitutes the core of the Arabic plagiarism detection system, with its n-gram segmentation and metadata annotation, and morphological annotation of each word in the collection. The corpus is accessed through our plagiarism detection system as we will explain in Section 4. Preprocessing, as explained in Fig. 2, includes removal of diacritics, punctuation and special characters. Letterform unification (i.e. "إ،أ،آ" are normalized to "ا"), n-gram segmentation ($n = 1–7$), part-of-speech tagging, stemming, and tokenization are also performed at this stage. Table 5 shows the final distribution of the collected texts (i.e., 2312 dissertations) as per the Dewey categories. The corpus statistics will be outlined subsequently.

Experiments and discussion

Experimenting with the JUPlag corpus: analysis and statistics

As stated earlier, the goal of constructing the JUPlag corpus is twofold. First, it is intended to be used to detect plagiarism in students' assignments, reports, and new dissertations prior to submission for defense. Secondly, its unique design structure

[5] https://camel.abudhabi.nyu.edu/madamira/

Table 5 Detailed classification of the corpus texts based on Dewey's categories

Dewey (Level-2)	Arabic Categories	English Categories	Number of Dissertations
004	معالجة البيانات, علم الحاسوب	Data processing & computer science	17
020	علم المكتبات والمعلومات	Library & information sciences	
150	علم النفس	Psychology	10
210	الإسلام و علومه	Philosophy & theory of religion	397
220	القرآن الكريم وعلومه	Holy Quran and its Sciences	
290	الديانات الأخرى	Other religions	
301	علم الاجتماع والانثروبولوجيا	Sociology & anthropology	1338
302	التفاعل الاجتماعي	Social interaction	
303	العمليات الاجتماعية	Social processes	
304	العوامل المؤثرة في السلوك الاجتماعي	Factors affecting social behavior	
305	الجماعات الاجتماعية	Groups of people	
306	الثقافة ومؤسساتها	Culture & institutions	
307	المجتمعات	Communities	
320	العلوم السياسية	Political science (Politics & government)	
330	الاقتصاد	Economics	
340	القانون	Law	
350	الإدارة العامة	Public administration & military science	
360	الخدمات الاجتماعية؛ الجمعيات	Social problems & services; associations	
370	التربية والتعليم	Manners & education	
380	التجارة , الاتصالات, النقل	Commerce, communications, transportation	
401	اللغويات, علم اللغة	Philosophy & theory; international languages	42
410	اللغة العربية	Linguistics	
507	التعليم والموضوعات ذات الصلة	Education, research, related topics	17
510	الرياضيات	Mathematics	
530	الفيزياء	Physics	
540	الكيمياء والعلوم ذات الصلة	Chemistry & allied sciences	
550	علوم الأرض (الجيولوجيا)	Earth sciences	
570	علوم الحياة (علم الأحياء)	Biology	
580	علوم النبات	Plants	
610	العلوم الطبية (الطب)	Medicine & health	103
630	الزراعة	Agriculture & related technologies	
650	إدارة الأعمال والخدمات المساعدة	Management & auxiliary services	
710	تخطيط المدن والعمران	Area planning & landscape architecture	147
720	الهندسة المعمارية (العمارة)	Architecture	
780	الموسيقى	Music	
790	الفنون الترفيهية والاستعراضية	Recreational & performing arts	
808	البلاغة الأدبية ومجموعات الأدب	Rhetoric & collections of literary texts	134
809	تاريخ ونقد الأدب	History, description & critical appraisal	
810	الأدب العربي	Arabic literature	
907	التعليم والبحث والموضوعات ذات الصلة	Education, research, related topics of history	107
910	الجغرافيا والرحلات	Geography & travel	
930	تاريخ العالم القديم	History of ancient world to ca. 499	
940	تاريخ أوروبا العام	History of Europe	

Table 5 Detailed classification of the corpus texts based on Dewey's categories *(Continued)*

Dewey (Level-2)	Arabic Categories	English Categories	Number of Dissertations
950	تاريخ آسيا العام الشرق الأقصى	History of Asia	
Total			2312

provides a knowledgebase for linguistic analysis, language teaching, and the learning of plagiarism-free writing. In this respect, the user can query a subset of the corpus to retrieve language patterns that are favored in the particular discipline to which this subcorpus is dedicated. For example, frequency word lists can be generated for a particular discipline; thus, technical lexicography can be facilitated. The corpus can also be used to demonstrate plagiarism-avoidance strategies in research methodology courses and to teach linguistic patterns in writing and linguistic analysis courses.

To experiment with this corpus, a linguistic concordancer described in Hammo et al. (2016) is used to inquire about words and n-gram sentences in the database. Metadata such as subject topic, author name, and publication date are used to facilitate search and filter retrieved data.

Word statistics

The JUPlag academic corpus has around 60 million words and (825,363) word types. Table 6 shows the top 20 words in the corpus, their English translation and their frequencies.

It is interesting to observe that the most frequent words in this predominantly social science corpus are general academic words and none of them is discipline-specific. Probably, the only words that betray the nature of the texts in this corpus are the words for 'God', 'Mohammad', and 'Jordan' since the theses/dissertations were produced in a Muslim country, Jordan.

It is also interesting to have an insight into the content of the corpus from a statistical perspective. In this context, "information theory states that messages maximize their capacity to convey information when the content follows Zipf's law". For a text corpus, Zipf's law specifies that, given a large sample of words, if w_1 is the most common word in the corpus, w_2 is the next most common, then the frequency of the i^{th}

Table 6 Top 20 frequent words in the corpus and their English translation

Rank	Arabic Word	English Meaning	Frequency	Rank	Arabic Word	English Meaning	Frequency
1	الدراسه	The study	260,127	11	دراسه	Study	76,374
2	الله	Allah (God)	222,662	12	وجود	Existence	75,484
3	الطلبه	Students	103,332	13	رقم	Figure	75,422
4	محمد	Mohammad	94,735	14	استخدام	Use	69,369
5	العمل	Work	94,469	15	يمكن	Possible	68,652
6	الأول	First	94,151	16	عام	Year	67,907
7	العربيه	Arabic	93,333	17	عدم	Un/Not	66,971
8	الثاني	Second	87,322	18	الأردن	Jordan	66,684
9	مستوى	Level	81,435	19	نتائج	Results	66,390
10	دار	House	76,968	20	بشكل	Form	65,314

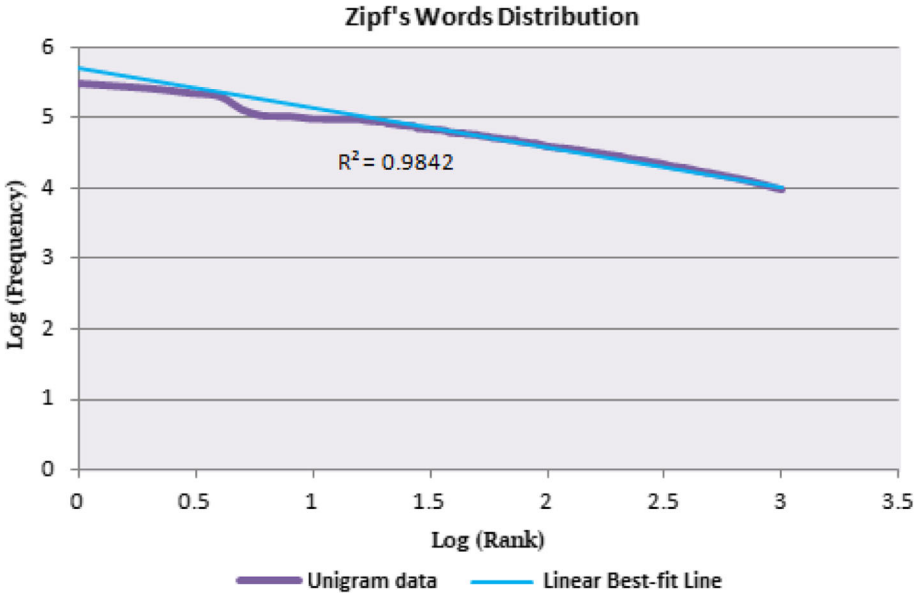

Fig. 3 Zipf's distribution of the top 1000 words in the JUPLag corpus

most common word is inversely proportional to its rank in the frequency table. So word number i has a frequency proportional to $1/i$.

To visualize how words are distributed across the corpus, we used a log-log scatter chart which plots the collection's frequency of a word as a function of its rank for the top 1000 words in the JUPlag corpus as shown in Fig. 3. The linear trendline shown along the curve in the chart is a best-fit straight line that is used with simple linear datasets to determine if the data follows Zipf's law. It is most reliable when the calculated R-squared (R^2) value of the best-fit line is equal or close to 1. For the unigram

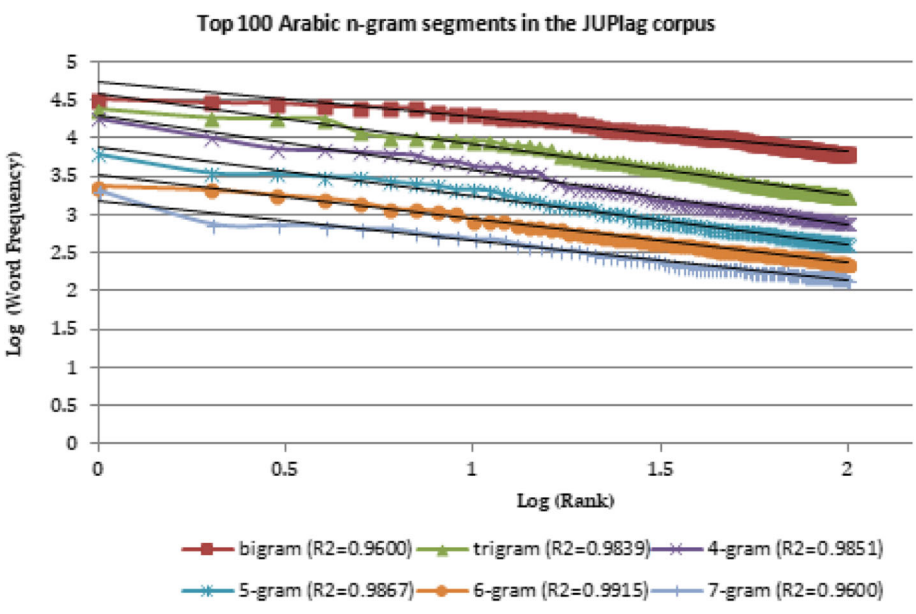

Fig. 4 Zipf's distribution of the top 100 Arabic n-gram segments in JUPLag corpus

sample, the R^2 value was (0.9842), which indicates that the unigram distribution is around the Zipf's law distribution.

Sentence statistics

According to Coxhead (2000), Zipf's law has been used often by language educators to identify the most common words/sentences for purposes of teaching foreign languages. Figure 4 shows the log-log scatter chart plot of the top 100 n-gram segments; it depicts how distribution of the top 100 n-gram chunks in JUPlag observes Zipf's law. Figure 4 also shows that the R^2 values for all trendlines corresponding to the n-gram segments are very close to 1, which again indicates excellent fit of the n-gram segments to Zipf's law distribution.

Now let's take a look at the top 10 n-gram segments sampled from the JUPlag corpus as shown in Table 7.

From Table 7, it is interesting to observe that the most frequent n-gram segments in this predominantly social science corpus are statistical expressions. Examples of bigrams are "عينة الدراسة" (study sample) and "المتوسطات الحسابية" (statistical means). An

Table 7 Top 10 frequent n-gram segments in the JUPlag corpus

Rank	*N* = 2	*N* = 3	*N* = 4	*N* = 5	*N* = 6	*N* = 7
1	33,453	24,438	18,120	6203	2378	2078
	عينة الدراسة	صلى الله وسلم	المتوسطات الحسابية والانحرافات المعيارية	رسول الله صلى الله وسلم	رسالة ماجستير منشورة الجامعة الأردنية عمان	رسالة ماجستير منشورة الجامعة الأردنية عمان الأردن
2	29,861	18,789	10,226	3520	2117	763
	المتوسطات الحسابية	الحسابية والانحرافات المعيارية	وجود فروق دلالة إحصائية	عدم وجود فروق دلالة احصائية	ماجستير منشورة الجامعة الأردنية عمان الأردن	منشورة جامعة عمان العربية للدراسات العليا عمان
3	29,468	18,330	7266	3410	1838	731
	دلالة احصائية	المتوسطات الحسابية والانحرافات	النبي صلى الله وسلم	فروق دلالة احصائية مستوى الدلالة	تم حساب المتوسطات الحسابية والانحرافات المعيارية	جامعة عمان العربية للدراسات العليا عمان الأردن
4	26,967	18,086	6979	3135	1660	717
	صلى الله	فروق دلالة احصائية	رسول الله صلى الله	منشورة الجامعة الأردنية عمان الأردن	وجود فروق دلالة احصائية مستوى الدلالة	وجود فروق دلالة احصائية مستوى الدلالة الفا
5	26,167	11,953	6563	3092	1469	660
	محمد بن	وجود فروق دلالة	الله صلى الله وسلم	وجود فروق دلالة احصائية مستوى	فروق دلالة احصائية مستوى الدلالة الفا	فروق دلالة احصائية مستوى الدلالة الفا اقل
6	25,698	10,390	6262	2733	1224	658
	الله وسلم	المملكة العربية السعودية	فروق دلالة احصائية مستوى	رسالة ماجستير منشورة الجامعة الأردنية	استخراج المتوسطات الحسابية والانحرافات المعيارية تمّ	عدم وجود فروق دلالة احصائية مستوى الدلالة
7	25,373	10,127	6023	2528	1179	590
	عبد الله	دلالة احصائية مستوى	دار الكتب العلمية بيروت	الجدول المتوسطات الحسابية والانحرافات المعيارية	توجد فروق دلالة احصائية مستوى الدلالة	هل توجد فروق دلالة احصائية مستوى الدلالة
8	22,130	9707	5001	2419	1128	535
	المتوسط الحسابي	افراد عينة الدراسة	دلالة احصائية مستوى الدلالة	ماجستير منشورة الجامعة الأردنية عمان	عدم وجود فروق دلالة احصائية مستوى	توجد فروق دلالة احصائية مستوى الدلالة الفا
9	21,992	9463	4979	2145	1060	518
	نتائج الدراسة	وزارة التربية والتعليم	رسالة ماجستير منشورة جامعة	دلالة احصائية مستوى الدلالة الفا	هل توجد فروق دلالة احصائية مستوى	رسالة ماجستير منشورة جامعة اليرموك اربد الأردن
10	21,133	9325	4320	2131	842	482
	فروق دلالة	رسالة ماجستير منشورة	توجد فروق دلالة احصائية	توجد فروق دلالة احصائية مستوى	منشورة جامعة عمان العربية للدراسات العليا	المتوسطات الحسابية والانحرافات المعيارية لاستجابات افراد عينة

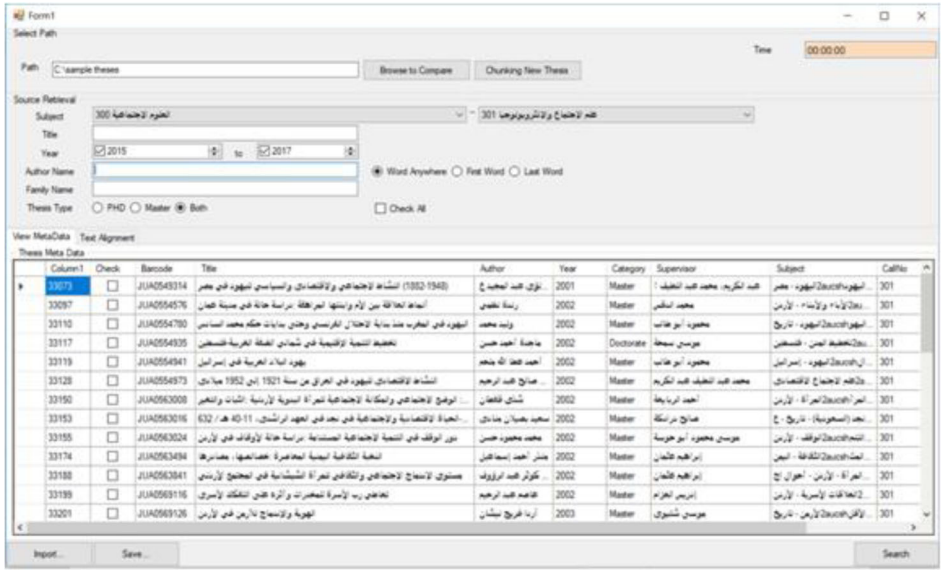

Fig. 5 Graphical user interface of the PD system

example of a trigram segment is "الحسابية والانحرافات المعيارية" (statistical and standard deviations), 4-g to 7-g samples are: "المتوسطات الحسابية والانحرافات المعيارية" (statistical means and standard deviations), "عدم وجود فروقات دلالة احصائية" (there are no statistically significant differences), "تم حساب المتوسطات الحسابية والانحرافات المعيارية" (statistical means and standard deviations were calculated) and "وجود فروق دلالة احصائية مستوى الدلالة الفا" (there are statistically significant differences level alpha) respectively.

Observation also indicates that most of the dissertations in the social sciences in this corpus appear to require surveys, collecting and analyzing data, and calculating statistics. Hence, the JUPlag corpus can be used as a knowledge base for the teaching of empirical research.

Experimenting with the plagiarism detection system

To experiment with the academic plagiarism corpus, we implemented a plagiarism detection (PD) system as shown in Fig. 5. The PD system is guided by the DDC system and the JU Library's metadata for retrieval and discovery of plagiarism. A new submitted dissertation can be checked for plagiarism either in a specific Dewey category

Table 8 Characteristics of the test dataset

Segment	Untampered Test Dataset		Test Dataset with Plagiarized Paragraphs		Test Dataset with Plagiarized Sentences	
	Count	unique count	Count	unique count	count	Unique count
unigram	632	413	735	487	678	441
bigram	631	586	734	682	677	626
trigram	630	618	733	718	676	662
4-g	629	624	732	725	675	672
5-g	628	627	732	730	674	673
6-g	627	627	731	731	673	673
7-g	626	626	729	729	672	672

Table 9 Samples of unique bigrams labeled as plagiarized

Title of Source Dissertation	Plagiarized Bigrams	English Translation	Frequency
النظرية البنائية الوظيفية والتركيز على إسهامات روبرت ميرتون	الاجتماعية والاقتصادية	Social & economical	15
	الاجتماعية والثقافية	Social & cultural	7
عمل الزوجة وأثره على أوضاعها الأسرية: دراسة ميدانية على عينة في مدينة مسقط	اوضحت نتائج	Results showed	2
	الاجتماعية والاقتصادية	Social & economical	9

(subcorpus) or it can be checked against the entire JUPlag corpus. Our experimentation here utilizes both types.

To test the PD system, we obtained a new dissertation from the School of Graduate Studies at JU. The dissertation was in the field of "Sociology" "علم الاجتماع" which by Dewey's classification belongs to the superordinate class of "Sociology and Anthropology" "علم الاجتماع والأنثروبولوجيا" subclass. At this early stage in our project, we only focused on copy-&-paste phenomena, verbatim plagiarism. The test dataset consists of three pages that were extracted from this new dissertation. We created two datasets: One was injected with two plagiarized paragraphs; the other was injected with multi-instances of plagiarized sentences. Both datasets went through preprocessing and segmentation into n-grams of strings as discussed in the previous section. The value of n has been set to 2–7 g. Table 8 shows the characteristics of the three datasets in the untampered form, with plagiarized paragraphs, and with plagiarized sentences. The *count* column lists the frequency of occurrence of the n-gram segments, the *unique count* column lists the frequency of such segments when repeated sequences are excluded.

Experiment I: plagiarism detection in the original dataset

The first experiment ran the plagiarism detection system through the untampered test dataset in six iterations of segmentation: 2-gram, 3-gram, 4-gram, 5-gram, 6-gram, and 7-gram segmentation. It ran it against the "Sociology and Anthropology" subcorpus (cf. Table 8). The success rate of plagiarism detection for a dissertation (D) is calculated by Eq. 1.

$$Reported\ Plag_D = \frac{detected\ plagiarized\ unique\ n\text{–}grams\ in\ D}{all\ unique\ n\text{–}grams\ in\ D} \times 100\% \tag{1}$$

The PD system labeled as 'plagiarized' (256) out of the (586) bigrams in the untampered test dataset (i.e., 43.68%) (cf. Table 8). Table 9 shows samples of the bigram segments that were labeled as 'plagiarized'. The first column lists the titles of the source

Table 10 Detected trigram segments

Title of Dissertation	Detected Trigrams	English Translation	Frequency
النظرية البنائية الوظيفية والتركيز على إسهامات روبرت ميرتون	السياسية والاقتصادية والاجتماعية	Political, economical and social	3
مشكلات المرأة الصحفية العاملة في الصحف اليومية الأردنية	السياسية والاقتصادية والاجتماعية	Political, economical and social	1
أثر المتغيرات الاقتصادية والاجتماعية على الاتجاهات السياسية لأعضاء هيئة التدريس في الجامعة الأردنية	الظروف الاقتصادية والاجتماعية	Economical and social conditions	1
عمل الزوجة وأثره على أوضاعها الأسرية :دراسة ميدانية على عينة في مدينة مسقط	جاءت العوامل الاقتصادية	Economical factors were	1

Table 11 Detected trigrams in the context

Detected Trigrams	Source in the Subcorpus	Title of Dissertation
السياسية والاقتصادية والاجتماعية	...مكونات الحياة **السياسية والاقتصادية والاجتماعية** وهذا يتطلب مزيداً من التخصص...	النظرية البنائية الوظيفية والتركيز على إسهامات روبرت ميرتون
السياسية والاقتصادية والاجتماعية	...والعقبات **السياسية والاقتصادية والاجتماعية** التي تواجه الإعلاميات العربيات...	مشكلات المرأة الصحفية العاملة في الصحف اليومية الأردنية
الاقتصادية الظروف والاجتماعية	...**الظروف الاقتصادية والاجتماعية** الحالية والتصدي لها قبل الاتجاه بجرأة...	أثر المتغيرات الاقتصادية والاجتماعية على الاتجاهات السياسية لأعضاء هيئة التدريس في الجامعة الأردنية
جاءت العوامل الاقتصادية	...وتأكيد الذات واكتساب الخبرة والاحتكاك بالمجتمع وبالتالي **جاءت العوامل الاقتصادية** مقدمة الى الدوافع...	عمل الزوجة وأثره على أوضاعها الأسرية: دراسة ميدانية على عينة في مدينة مسقط

Table 12 Detected 4-grams in the context

Detected 4-g	Source in the Subcorpus	Title of Dissertation
شهدت السنوات الأخيرة تزايدا	**شهدت السنوات الأخيرة تزايدا** في معدلات توظيف المرأة بشكل كبير في مختلف المستويات التعليمية...	عمل الزوجة وأثره على أوضاعها الأسرية: دراسة ميدانية على عينة في مدينة مسقط
السياسية والاقتصادية والاجتماعية والثقافية	ولكن التطورات والتغيرات أصابت المجتمعات في جميع الجوانب **السياسية والاقتصادية والاجتماعية والثقافية** ورافقه تغير وجهة النظر السابقة...	النظرية البنائية الوظيفية والتركيز على إسهامات روبرت ميرتون

Table 13 Results of experiment I: plagiarism-labeling in the untampered test dataset

N-gram Segments	Retrieved Dissertations	Segments in the Test Dataset	Segments Identified as Plagiarized	Reported Plagiarism Ratio
3	4	618	15	2.43%
4	2	624	2	0.32%
5	0	627	0	0.00%
6	0	627	0	0.00%
7	0	626	0	0.00%

dissertations where the detected bigrams were found, the second lists the detected bigrams, and the last lists the frequency of occurrence of these bigrams in the respective dissertations.

Bigram matching, however, is of little significance as bigrams hardly ever express a complete thought. It is not unexpected for matches to be found between bigrams in different dissertations since most two-word strings hold general concepts. Therefore, bigram matches might not be indicative of direct verbatim plagiarism.

When the PD system ran through the trigram segments, it labeled (15) out of the (618) trigrams in the test dataset as instances of plagiarism, i.e., the reported plagiarism rate was 2.43% (cf. Table 8). They were found in four dissertations. Table 10 shows a sample of the detected trigram segments.

A closer look at the detected trigrams shows that they also denote general concepts (see Table 11). However, many scholars consider the similarity of n-gram segments of four or more consecutive words to be verbatim plagiarism and hence it must be labeled as such. For example, Hexham (2005) treated the similarity of strings of four consecutive words as plagiarism, Roig (1999) five words, and Sorokina, Gehrke, Warner, and Ginsparg (2006) seven words.

When the PD system ran through the 4-gram iteration of the test dataset, it labeled only (2) out of the (624) 4-gram segments as instances of plagiarism, i.e., the reported plagiarism rate was 0.32% (cf. Table 8). Table 12 shows the detected 4-gram plagiarism.

Again, the 4-gram segments express general concepts and they hardly constitute genuine plagiarism. Although, 5-gram strings according to Roig (1999) are considered a good starting point for potential plagiarism, in this first experiment we could not find in the "Sociology and Anthropology" subcorpus any suspicious segments of five, six or seven consecutive words. Table 13 summarizes the results of the first experiment.

This experiment has demonstrated that when there is no intended plagiarism, a PD system can still label short segments as 'plagiarized'; the shorter the segment is, the more susceptible it is to misidentification as an instance of plagiarism. Passing a verdict of 'plagiarized segments' should be left to the discretion of the human. The machine

Table 14 Plagiarism-simulated paragraphs injected in the original dataset

Paragraph-1 inserted on page1	شهدت السنوات الأخيرة تزايداً في معدلات توظيف المرأة بشكل كبير في مختلف المستويات التعليمية كنتيجة طبيعية لمخرجات المؤسسات التعليمية المختلفة كجامعة السلطان قابوس وكليات التربية ومعاهد العلوم الصحية. وبلغ عدد الموظفات العمانيات المعينات العام م حوالي بنسبة إجمالي الموظفين العمانيين المعينين.
Paragraph-2 inserted on page2	ولكن التطورات والتغيرات أصابت المجتمعات في جميع الجوانب السياسية والاقتصادية والاجتماعية والثقافية ورافقها تغير وجهة النظر السابقة المرتبطة بالأدوار المناطة بكلا الجنسين الرجل المرأة وتحسن مستوى تعليم المرأة وخرجت للعمل وقامت المؤسسات مثل دور الرعاية وتنشئة الأطفال بالإضافة الى الأسرة بفتح المجال أمام المرأة بأن تأخذ أدواراً جديدة في المجتمع؛ لذلك ينظر الوظيفيون الى الأمر بأنه يتطلب نوعاً من التعديل في النظم الاجتماعية السابقة من أجل عدم حصول توتر وصراع داخل المجتمع بسبب اختلاف تقسيم الأدوار.

Table 15 Results of the paragraph simulated-plagiarism experiment

N-gram Segments	Retrieved Dissertations	Segments in the Test Dataset	Segments with Simulated Plagiarism	Segments Identified as Plagiarized	Reported Plagiarism Ratio
3	4	718	99	114	15.88%
4	2	725	97	99	13.93%
5	2	730	95	95	13.01%
6	2	731	93	93	12.72%
7	2	729	91	91	12.48%

can only point to the similarity it identified. Causes of this similarity, however, might be totally unrelated to plagiarism as demonstrated by the bigram and trigram detection.

Experiment II: detecting paragraph simulated-plagiarism

In the second experiment, the original test dataset was injected with two paragraphs extracted randomly from the "Sociology and Anthropology" subcorpus to simulate an act of plagiarism. The two paragraphs, shown in Table 14, were inserted into the first and second pages of the original test dataset. For the characteristics of the dataset with paragraph simulated-plagiarism see Table 8.

As established by the first experiment, bigram segments are too general to be considered as direct plagiarism. Hence, we ran the PD system through the test dataset with the two plagiarized paragraphs after segmenting it into the 3–7 g iterations.

The results of the second experiment are given in Table 15. The third column lists the number of segments after the insertion of the segments from the two plagiarism-simulated paragraphs (cf. Table 8). The fourth lists the number of n-gram segments that the plagiarized paragraphs consist of. The fifth lists the number of segments that the PD system labeled as 'plagiarized'. Notice that the values in the fifth column are higher than those in the fourth. The reason is that the PD system was able to detect all the simulated plagiarism and added the number of segments it had labeled as 'plagiarized' in the untampered dataset. For instance, the PD system labeled 114 as 'plagiarized' trigrams, 99 of which are trigrams in the plagiarism-simulating paragraphs and 15 trigrams labeled as 'plagiarized' in the original test dataset as explained in experiment I.

Table 16 Plagiarism-simulated sentences as injected in the original test dataset

Plagiarism-simulated sentences	Page#	Paragraph#
المتوسطات الحسابية والانحرافات المعيارية للكفاءة المعلوماتية لمكتبات	1	1
الأنماط البنائية وحالاتها	1	2
والانحرافات المعيارية للكفاءة التعليمية لمكتبات المدارس الأساسية	1	3
والبالغ عددهم وطلبة المرحلة الأساسية العليا الصفوف	1	4
أسئلة الدراسة استخراج التكرارات والنسب	2	1
تناول الظاهرة موضوع	2	3
منطقية يجري بمقتضاها	2	4
بصائر أحوال الظاهرة النحوية	3	1
استخراج التكرارات والنسب المئوية والمتوسطات الحسابية والانحرافات	3	2
أهمية التحليل الحركي	3	3

Table 17 Results of experiment III where the dataset is injected with plagiarism-simulated sentences and run against the subcorpus

N-gram Segments	Retrieved Dissertations	Segments in the Test Dataset	Segments with Simulated Plagiarism	Segments Identified as Plagiarized	Reported Plagiarism Ratio
3	4	662	28	15	2.27%
4	2	672	19	2	0.30%
5	0	673	13	0	0.00%
6	0	673	8	0	0.00%
7	0	672	4	0	0.00%

Experiment III: detecting plagiarism-simulated sentences injected in the dataset

In the third experiment, the original test dataset was injected with ten plagiarism-simulated sentences that were extracted randomly from the JUPlag corpus at large, rather than the "Sociology and Anthropology" subcorpus as the case was in the second experiment. The rationale was that we wanted to verify how our PD system would behave when the source of plagiarism is outside the scope of its corpus.

The ten plagiarized sentences are of variable word counts, 3 to 7 grams in length. They were appended to the original dataset in different paragraphs, with some injected on the first page, some on the second, and some on the third as shown in Table 16. For the characteristics of the new test dataset with plagiarism-simulated sentences, see Table 8. The PD system ran through this test dataset against the Sociology and Anthropology subcorpus.

A summary of the results of this experiment are in Table 17, where column 3 has the number of segments after insertion of the ten plagiarism-simulated sentences (cf. Table 8). In the next column are the number of n-grams that the plagiarized sentences consist of. Our PD system reports, in the last column, the plagiarism ratio as calculated by Eq. (1).

The table shows the PD system to have failed to detect any of the plagiarized n-gram segments of the sentences that were injected in the test dataset. The system, however, continued to label 15 trigrams and two of the 4-gram segments as 'plagiarized'. This is reminiscent of experiment I. This demonstrates that plagiarism from sources not covered by the PD corpus is likely to pass undetected.

To verify the efficiency of our PD system when the plagiarism lies within the scope of its corpus but without particularization of topic, the same experiment was run again but this time against the entire JUPlag corpus. It demonstrated that the system was perfectly capable of spotting plagiarized sentences even when the topic is not specified,

Table 18 Results of experiment III where the dataset is injected with plagiarism-simulated sentences and run against the entire corpus

N-gram Segments	Segments in Test Dataset	Segments with Simulated Plagiarism	Segments Identified as Plagiarized	Reported Plagiarism Ratio
3	662	28	159	24.02%
4	672	19	57	8.48%
5	673	13	20	2.97%
6	673	8	11	1.63%
7	672	4	5	0.74%

Table 19 Samples of plagiarized n-gram segments

N-gram Segments	Suspicious Plagiarism Sentences
5-gram	خلصت الدراسة مجموعة التوصيات ابرزها
	الدراسة مجموعة التوصيات ابرزها ضرورة
	توصلت الدراسة مجموعة النتائج اهمها
	شتى مجالات الحياة السياسية والاقتصادية
6-gram	اسئلة الدراسة تم استخراج التكرارات والنسب
	الدراسة تم استخراج التكرارات والنسب المئوية
7-gram	المئوية اسئلة الدراسة تم استخراج التكرارات والنسب

provided that the plagiarized source is in its corpus. See Table 18 for a summary of results and Table 19 for a sample of identified plagiarism.

In addition, the PD system labeled more n-gram segments other than the ones reported in experiment I. For instance, the PD system labeled (159) plagiarized trigram segments in the test dataset. This number includes the plagiarism-simulated trigrams (28), the (15) trigrams segments labeled in the subcorpus from experiment I, in addition to (116) new trigrams segments detected in the entire JUPlag corpus.

Notice in Table 18 that the PD system identified exceedingly more than the injected trigrams and 4-gram segments, but beginning from 5-grams the plagiarism yield became more reasonable. This goes to support Roig's (1999) definition of plagiarism as "the appropriation of strings of five consecutive words or longer. (p.973)" since shorter n-gram segments hardly ever constitute propositions. Even with 5-, 6-, and 7-gram segments, the system overestimated plagiarism by seven, three, and one segment respectively. This distortion indicates that the longer the segment is, the more confident the identification.

Conclusion and research directions

We presented above a plagiarism detection corpus built for Arabic and designed especially for academic purposes. JUPlag is organized in accordance with the Dewey classification system and is guided by the metadata adopted by the Library of the University of Jordan. Although this corpus is still under construction, research on Arabic that is carried out by the international community may benefit from it. It can use it in its current state for the detection of plagiarism in Arabic dissertations and articles prior to final submission. It can also be beneficial for the development of new plagiarism detection tools. It may also be used for corpus-based and corpus-driven linguistic analyses, for language learning and teaching, for lexicography, and for teaching research methodology. We showed here the stages of corpus construction and the challenges encountered. To test the reliability of the corpus and PD system, we conducted a set of experiments with multi-instances of plagiarism-simulated paragraphs and sentences deliberately injected in a test dataset. Experimental results proved both the corpus and the system to be quite efficient in detecting n-gram verbatim plagiarism. It has been demonstrated here that it is indispensable for an extrinsic plagiarism detection system to have an authentic, big, versatile, properly classified and richly annotated reference corpus. It has also been confirmed that verbatim plagiarism detection is only reliable when the similarity-matching unit is longer than 4-g. In the next phase of this project, the reference

corpus will be expanded to encompass all the dissertations in the Thesis Repository of the Union of Arab Universities. The PD system will also utilize a variety of plagiarism detection techniques. Future research may focus on the expansion and representativeness of the corpus; the bigger the corpus and the more representative it is of disciplines in the humanities and social sciences, the more efficient plagiarism detection will be. Detection may also be complemented by the use of intrinsic, machine learning, and deep machine learning techniques.

Abbreviations

2 L-APD: Two-Level Plagiarism Detection System; AraPlagDet: Arabic Plagiarism Detection; ASCII: American Standard Code for Information Interchange; BKDR: Brian Kernighan and Dennis Ritchie; CNN: Convolutional Neural Network; DDC: Dewey decimal classification; EPD: Extrinsic plagiarism detection; ExAraPlagDet: External Arabic Plagiarism Detection; HYPLAG: Hybrid Plagiarism; IPD: Intrinsic plagiarism detection; IR: Information Retrieval; JU: The University of Jordan; JUPlag: Jordan University Plagiarism corpus; LCS: Longest Common Substring; MADAMIRA: Morphological Analysis and Disambiguation of Arabic; NLP: Natural Language Processing; OSAC: Open Source Arabic Corpus; PAN@FIRE: PAN-Forum for Information Retrieval Evaluation; PD: Plagiarism detection; PDS: Plagiarism detection software; POST: Part-of-speech tagger; TF*IDF: Term Frequency-Inverse Document Frequency

Acknowledgments

The authors would like to thank the director of the JU library, Dr. Nashrawan Al-Tahat and her IT staff for giving us restricted access to the dissertation repository and for providing the necessary computing facilities for the development and analysis of the JUPlag corpus and for permission to experiment with our PD system.

Authors' contributions

This research is part of a dissertation research conducted by the first author, EA-T. All authors read and approved the final manuscript.

Funding

Data collection is funded by the JU Library as part of the doctoral program collaboration.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to JU copyright policy but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Computer Science Department, King Abdullah II School of Information Technology, University of Jordan, Amman, Jordan. [2]Computer Information Systems Department, King Abdullah II School of Information Technology, University of Jordan, Amman, Jordan. [3]Department of Foreign Languages, College of Arts, University of Sharjah, Sharjah, United Arab Emirates.

Received: 17 May 2019 Accepted: 20 November 2019
Published online: 16 January 2020

References

Abdelrahman, Y. A., Khalid, A., & Osman, I. M. (2017). A method for Arabic documents plagiarism detection. *International Journal of Computer Science and Information Security*, *15*(2), 79.

AlSallal, M., Iqbal, R., Palade, V., Amin, S., & Chang, V. (2019). An integrated approach for intrinsic plagiarism detection. *Future Generation Computer Systems, 96*, 700–712.

Baba, K., Nakatoh, T., & Minami, T. (2017). Vector representation of words for plagiarism detection based on string matching. In *International conference on human interface and the management of information*, (pp. 341–350). Cham: Springer.

Bensalem, I., Boukhalfa, I., Rosso, P., Abouenour, L., Darwish, K., & Chikhi, S. (2015). Overview of the AraPlagDet PAN@ FIRE2015 shared task on Arabic plagiarism detection. In *FIRE workshops*, (pp. 111–122).

Beute, N., Van Aswegen, E. S., & Winberg, C. (2008). Avoiding plagiarism in contexts of development and change. *IEEE Transactions on Education*, *51*(2), 201–205.

Bolkan, J. V. (2006). Avoid the plague: Tips and tricks for preventing and detecting plagiarism. *Journal of Vocational Education and Training*, *33*, 4–14.

Chan, L. M., Comaroni, J. P., Mitchell, J. S., & Satija, M. P. (1996). *Dewey decimal classification: A practical guide.* Albany: Forest Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

Devlin, M., & Gray, K. (2007). In their own words: A qualitative study of the reasons Australian university students plagiarize. *High Education Research & Development*, *26*(2), 181–198.

DeVoss, D., & Rosati, A. C. (2002). It wasn't me, was it? Plagiarism and the Web. *Computers and Composition*, *19*(2), 191–203.

Elhadi, M., & Al-Tobi, A. (2008). Use of text syntactical structures in detection of document duplicates. In *Third international conference on digital information management*, (pp. 520–525). London, UK.

Eret, E., & Ok, A. (2014). Internet plagiarism in higher education: Tendencies, triggering factors and reasons among teacher candidates. *Assessment & Evaluation in Higher Education*, *39*(8), 1002–1016.

Fister, B. (2009). The Dewey dilemma. *Library Journal*, *134*(16), 22–25.

Franklin-Stokes, A., & Newstead, S. (1995). Undergraduate cheating: Who does what and why? *Studies in Higher Education*, *20*(2), 159–172.

Frye, B. L. (2016). Plagiarism is not a crime. *Duquesne University Law Review*, *54*, 133 https://uknowledge.uky.edu/law_facpub/532. Visited 30 Mar 2019.

Ghanem, B., Arafeh, L., Rosso, P., & Sánchez-Vega, F. (2018). HYPLAG: Hybrid Arabic text plagiarism detection system. In *International conference on applications of natural language to information systems*, (pp. 315–323). Cham: Springer.

Gipp, B., & Beel, J. (2010). Citation based plagiarism detection: A new approach to identify plagiarized work language independently. In *Proceedings of the 21st ACM conference on hypertext and hypermedia*, (pp. 273–274). Toronto, Canada.

Gipp, B., & Meuschke, N. (2011). Citation pattern matching algorithms for citation-based plagiarism detection: Greedy citation tiling, citation chunking and longest common citation sequence. In *Proceedings of the 11th ACM symposium on Document engineering*, (pp. 249–258). Mountain View, California, USA.

Golub, K., Lykke, M., & Tudhope, D. (2014). Enhancing social tagging with automated keywords from the Dewey decimal classification. *Journal of Documentation*, *70*(5), 801–828.

Hammo, B., Yagi, S., Ismail, O., & AbuShariah, M. (2016). Exploring and exploiting a historical corpus for Arabic. *Language Resources and Evaluation*, *50*(4), 839–861.

Hexham, I. (2005). *Academic plagiarism defined*. Department of Religious Studies, University of Calgary http://people.ucalgary.ca/~hexham/content/articles/plague-of-plagiarism.html. Visited 30 Mar 2019.

Jenkins, C., Jackson, M., Burden, P., & Wallis, J. (1998). Automatic classification of web resources using Java and Dewey decimal classification. *Computer Networks and ISDN Systems*, *30*(1–7), 646–648.

Khoja, S., & Garside, R. (1999). *Stemming Arabic text*. Lancaster, UK, Computing Department, Lancaster University.

Khorsi, A., Cherroun, H., & Schwab, D. (2018). 2L-APD: A two-level plagiarism detection system for Arabic documents. *Cybernetics and Information Technologies*, *18*(1), 124–138.

Kong, L., Zhao, Z., Lu, Z., Qi, H., & Zhao, F. (2016). A method of plagiarism source retrieval and text alignment based on relevance ranking model. *International Journal of Database Theory and Application*, *9*(12), 35–44.

Leonardo, B., & Hansun, S. (2017). Text documents plagiarism detection using Rabin-Karp and Jaro-Winkler distance algorithms. *Indonesian Journal of Electrical Engineering and Computer Science*, *5*(2), 462–471.

Lukashenko, R., Graudina, V., & Grundspenkis, J. (2007). Computer-based plagiarism detection methods and tools: An overview. In *Proceedings of the 2007 international conference on computer systems and technologies*, (p. 40). Rousse, Bulgaria.

Mahmoud, A., Zrigui, A., & Zrigui, M. (2017). A text semantic similarity approach for Arabic paraphrase detection. In *International conference on computational linguistics and intelligent text processing*, (pp. 338–349). Cham: Springer.

Mahmoud, A., & Zrigui, M. (2017). Semantic similarity analysis for paraphrase identification in Arabic texts. In *Proceedings of the 31st Pacific Asia conference on language, information and computation*, (pp. 274–281).

Maurer, H. A., Kappe, F., & Zaka, B. (2006). Plagiarism-A survey. *Journal of Universal Computer Science*, *12*(8), 1050–1084.

Meuschke, N., Gipp, B., Breitinger, C., & Berkeley, U. (2012). CitePlag: A citation-based plagiarism detection system prototype. In *Proceedings of the 5th international plagiarism conference*. Newcastle upon Tyne, UK.

Meuschke, N., Siebeck, N., Schubotz, M., & Gipp, B. (2017). Analyzing semantic concept patterns to detect academic plagiarism. In *Proceedings of the 6th international workshop on mining scientific publications*, (pp. 46–53). Toronto, Canada.

Nakatoh, T., Baba, K., Yamada, Y., & Ikeda, D. (2011). Partial plagiarism detection using string matching with mismatches. In *International conference on informatics engineering and information science*, (pp. 265–272). Berlin: Springer.

Paul, M., & Jamal, S. (2015). An improved SRL based plagiarism detection technique using sentence ranking. *Procedia Computer Science*, *46*, 223–230.

Polydouri, A., Siolas, G., & Stafylopatis, A. (2017). Intrinsic plagiarism detection with feature-rich imbalanced dataset learning. In *International conference on engineering applications of neural networks*, (pp. 99–110). Cham: Springer.

Rahman, R. (2015). Information theoretical and statistical features for intrinsic plagiarism detection. In *Proceedings of the SIGDIAL 2015 conference, Prague, Czech Republic*, (pp. 144–148).

Roig, M. (1999). When college students' attempts at paraphrasing become instances of potential plagiarism. *Psychological Reports*, *84*(3), 973–982.

Ruipérez, G., & García-Cabrero, J.-C. (2016). Plagiarism and academic integrity in Germany. *Comunicar: Media Education Research Journal*, *24*(48), 9–17. https://doi.org/10.3916/C48-2016-01.

Saad, M. K., & Ashour, W. M. (2010). Osac: Open source arabic corpora. In *The 6th international conference on electrical and computer systems (EECS'10), Lefke, North Cyprus*.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*, 513–523.

Sapir, E., & Swiggers, P. (2008). *General linguistics*, (vol. 1). Walter de Gruyter. Berlin, New York.

Satija, M. P., & Martínez-Ávila, D. (2019). Plagiarism: An essay in terminology. *DESIDOC: Journal of Library & Information Technology*, *39*(2), 87–93. https://doi.org/10.14429/djlit.39.2.13937.

Si, A., Leong, H. V., & Lau, R. W. (1997). Check: A document plagiarism detection system. In *SAC*, (vol. 97, pp. 70–77).

Sorokina, D., Gehrke, J., Warner, S., & Ginsparg, P. (2006). Plagiarism detection in arXiv. In *Proceedings of the sixth international conference on data mining*, (pp. 1070–1075) Available at http://arxiv.org/abs/cs/0702012. Visited 30 Mar 2019.

Tschuggnall, M., & Specht, G. (2012). Plag-inn: Intrinsic plagiarism detection using grammar trees. In *International conference on application of natural language to information systems*, (pp. 284–289). Berlin: Springer.

Vani, K., & Gupta, D. (2017). Text plagiarism classification using syntax based linguistic features. *Expert Systems with Applications*, *88*, 448–464.

Velupillai, V. (2012). *An introduction to linguistic typology*. John Benjamins Publishing. Amsterdam, The Netherlands.

Vij, R., Soni, N. K., & Makhdumi, G. (2009). Encouraging academic honesty through anti-plagiarism software. In *Proceedings of the 7th international CALIBER, Pondicherry University*, (pp. 439–448).

Wise, M. J. (1996). YAP3: Improved detection of similarities in computer program and other texts. *ACM SIGCSE Bulletin*, *28*(1), 130–134 ACM.
Yang, Y. M. (1995). Noise reduction in a statistical approach to text categorization. In *Proceedings of SIGIR-95, 18th ACM international conference on research and development in information retrieval*, (pp. 256–263).
Zaher, M., Shehab, A., Elhoseny, M., & Osman, L. (2017). A new model for detecting similarity in Arabic documents. In *International conference on advanced intelligent systems and informatics*, (pp. 488–499). Cham: Springer.
Zu Eissen, S. M., & Stein, B. (2006). Intrinsic plagiarism detection. In *European conference on information retrieval*, (pp. 565–569). Berlin: Springer.

Publisher's Note