Implementation of a New Hybrid Method for Stemming of Arabic Text

Tahar Dilekh
Department of Computer Sciences
University of Hadj Lakhdar - Batna
Batna, ALGERIA

Ali Behloul
Department of Computer Sciences
University of Hadj Lakhdar - Batna
Batna, ALGERIA

ABSTRACT
In this paper, we propose a hybrid method that combines the application of three previously used techniques. These techniques deal with three key issues related to Arabic stemming including affix removal proposed by Kadri [1], dictionaries [2] and morphological analysis [3] [4] [5]. Thus, when solving these problems these techniques are applied individually and independently to solve associated stemming problems, which requires some adjustments to be implemented on each one of them.

Therefore, the main contribution of this experiment is to demonstrate the effectiveness of the hybrid method compared to other methods, and the choice of removing the suffix before prefix during the operation of Arabic stemming process.

General Terms
Natural language processing, Information retrieval.

Keywords
Information retrieval; Indexation; Tokenization; Stemming; Arabic language.

1. INTRODUCTION
Arabic is one of the six official languages of the United Nations and the mother tongue of more than 300 million people [6]. Recently, due to the growing number of internet users around the globe, information retrieval (IR) has become of great importance as an essential tool for all tasks of search on the web. The number of Arab Internet users in March, 2011 was about 77.8 millions which represents about 24.35 % of the population of the Arab world [6]. Relatively fewer Arabic search engines are currently available despite the enormous efforts put together to satisfy the needs of the growing number of Arabic internet users. Moreover, Arabic is a highly inflected language and has a complex morphological structure [7] [8] [9] [10], which makes information retrieval on Arabic texts require the basic form of the word (root or stem) [11]; therefore stemming process is necessary.

Arabic information retrieval has become increasingly important. This area of research has experienced a significantly great progress in recent decades. This latter is the main motivation for interest in studying the natural language processing.

Arabic has a very rich and complex morphology. It has an origin very different from European languages; it includes 28 letters and is written cursively from right to left. The morphological representation of Arabic is rather complex because of the morphological variation and the agglutination phenomenon.

In addition to the complex morphology, in written Arabic, the vowels (diacritics) are omitted and as a result of this omission, the words tend to have a higher level of ambiguity as well as the problem of the plural of irregular nouns (broken plural). In this case, a noun in plural takes another morphological form different from its initial form in singular.

To solve these problems and many others we base on the stemming algorithms or stemmers to group words based on semantic similarity. There are several types of stemming algorithms. The two most effective Arabic stemmers are Larkey's light stemmer [8] [12] and Khoja's [7] root-extraction stemmer.

We develop an information retrieval system dedicated to Arabic language based on a hybrid method in the phase of stemming that combines three known techniques: removal of affix proposed by Kadri [1], dictionaries [2] and morphological analysis [3] [4] [5].

The contribution of this paper is to provide a comprehensive analysis on a number of levels of information retrieval related issues particularly: (I) a study of the different methods of Stemming in Arabic, (II) applications of some methods stemming and (III) evaluating a performance of our hybrid method developed in this paper.

2. SYSTEM ARCHITECTURE
Our system can provide the test corpus with the possible index and allows the retrieval while optimizing costs in terms of time and storage space.

Schematically, the analysis involves the following phases:

- Unify the text encoding for either the corpus or for the queries;
- Normalize the corpus of texts and the queries;
- Split or separate the input text into sequences of lexical units (word);
- Eliminate the stop words;
- Determine for each word its morphological characteristics;
- Remove prefixes and suffixes based on morphological characteristics and various dictionaries;
- Determine the possible roots for each word based on dictionaries of patterns (AOUZANE) and roots;
- Weigh the terms generated;
- Create the index base ;

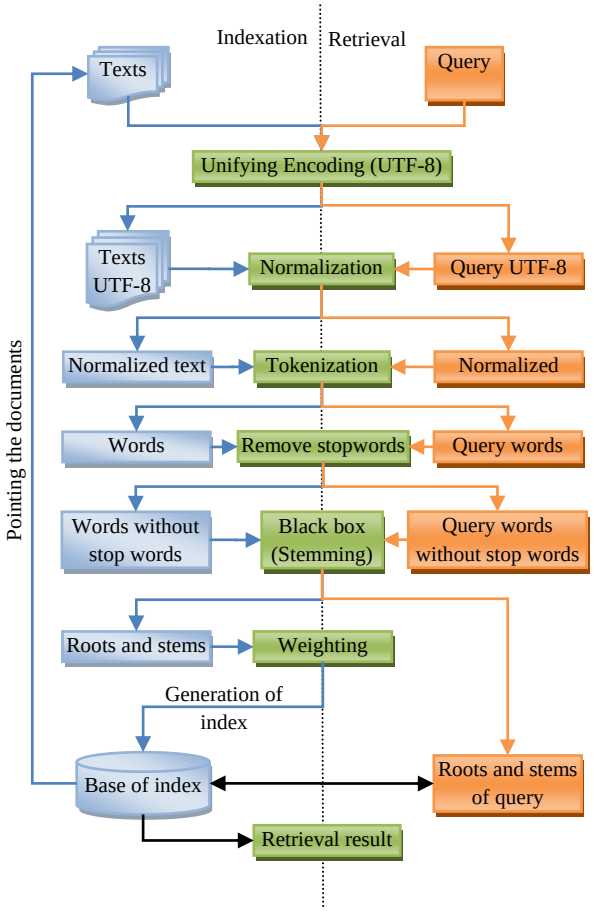

Fig 1: Architecture of our system.

Indexing (text input) and retrieval (user requests) are processed by these modules to obtain meaningful results, thus improving retrieval (Fig 1).

2.1 Encoding

The collection of texts and queries can be encoded differently, making them incomparable. For example: documents are represented in Unicode (UTF-8) and requests in ISO-8859-6 or any other encoding system. In order to standardize the documents with the queries, we must re-use converting tools between different encodings systems. Thus, everything would be converted into Unicode in our case.

2.2 Normalization

In our approach, normalization involves the following steps:

- Convert to Windows Arabic encoding (CP1256), if necessary;
- Remove punctuation;
- Remove diacritics (primarily weak vowels);
- Remove non-letters;
- Replace the إ or the أ initial by Alif nu ا;
- Replace the آ by the ا;
- Replace the ىء of order by the ئ;
- Replace the ى final by the ي;
- Replace the ة final by the ه.

The definitions of punctuation, diacritics and non-letters came from the stemmer of Khoja [7]. We improve normalization process significantly through two minor modifications. First, we replace the إ or أ by the ا bare Alif regardless of the position in the word. Second, we remove the Tatweel "ـ".

2.3 Tokenization

Tokenization is a necessary and meaningful step in natural language processing. The function of a tokenizer is to break down a text stream into segments so that they can be introduced into a morphological sensor or a position tagger. The tokenizer is responsible in defining boundaries of a word; it is based mainly on the white spaces and punctuation marks as delimiters between words or major segments (Fig 2).

Input	كل وعاء يضيق بما جعل فيه إلا وعاء العلمِ فإنه يتسع به. Each pot narrowing including making it, only a bowl of science; it expands by it.
Output	كل | وعاء | يضيق | بما | جعل | فيه | إلا | وعاء | العلم | فإنه | يتسع | به

Fig 2: Example of the Tokenizer.

2.4 Removing stop words

One major problem of indexing is to extract meaningful terms and to avoid stop words. There are two techniques to eliminate stop words:

- Using a list of stop words (also called anti-dictionary);
- The elimination of words over a certain number of occurrences in the collection.

We mainly use the first technique to remove stop words. In addition to that, we enrich the list of stop words by implementing the second technique. Applying the two combined techniques expands our list of stop words.

Although the removal of stop words has the advantage of reducing the number of indexing terms [13], it may reduce the recall rate: i.e. the proportion of relevant documents returned by the system to all relevant documents.

2.5 Stemming

The morphological process is the heart of Arabic information retrieval, in particular stemming. This latter plays an important role [14] [15]. We apply five different methods of stemming, and evaluate their performance; where each method integrates at least two techniques of stemming among three different techniques, including: morphological analysis, affix-removal and dictionaries, and then the method that performs well in information retrieval is adopted.

The Affix-Removal Technique uses a list of prefixes and suffixes to convert words to their base form (root or stem) [16] [17]. The morphological analysis technique is based on pattern matching to extract the root of the word [3] [7] [18]. This technique is used to decrease the number of improper removal cases of some affixes from words which have main letters that appear as affixes and to help convert the broken plural words to their singular forms. The dictionaries technique is adopted to resolve the improper removal of some prefixes and to handle the Arabicized words problem.

2.5.1 Stemming methods:

2.5.1.1 The method PS-M:

The method PS-M (Prefix then Suffix without Model/ Pattern) is the Kadri's light stemming [1].We adopt this Kadri's

method but with reduction of inflected word, by removing its prefixes, then its suffixes. At each step we check the existence of the resulting word in the dictionary of roots. If it exists, the process must be stopped; otherwise we repeat the process until we find one that dictionary of roots contain. When this process is done correctly, it becomes easy to extract the letters of the stem, for example "لنموه" (for growth) if its suffix is removed first "وه" (WIH), then we will lose the stem proper (Fig 3).

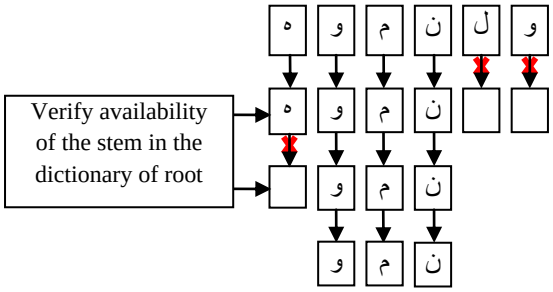

Fig 3: Example of the method PS-M.

2.5.1.2 The method SP-M:

The method SP-M (or Suffix Prefix without Pattern/Model) is the Kadri's light stemming [1]. We adopt this Kadri's method but with reduction of inflected word, by removing its suffixes, then its prefixes. For example the word "الالتزامات" (engagements), if we remove its first prefix "الا" (ALLI), we will lose the correct stem (Fig 4).

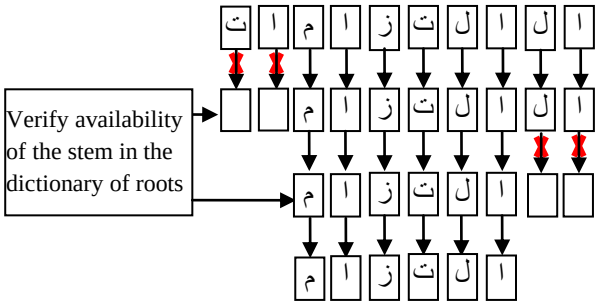

Fig 4: Example of the method SP-M.

2.5.1.3 The method PS+M (Prefix then Suffix with Model/ Pattern):

After removing all prefixes and suffixes of the inflected word, we compare it with all available patterns. If a pattern is found, we proceed to the extraction of letters which form the root; otherwise (if no pattern is found) we return the word as it is.

Removing some prefixes and suffixes of words helps reducing the number of patterns, facilitating the patterns matching process and allows multiple variations of the stem to be combined in the same pattern. As an example we did not keep any of these patterns: "استفعل" (ISTAFAALA), "مستفعل" (MOUSTAFAIL) because the two prefixes "است" (ISTA), "مست" (MOUSTA) are available. Instead, we remove all the prefixes and suffixes before comparing the word with its pattern. In this way we reduce the number of patterns and so finding the correct pattern.

We compare any word with patterns depending on their length, using a set of conditions to check the infix letters in the word. For example, the word "حواسيب" (Computers) has the length 6, so, first step, we look for patterns using the following condition:

Find a pattern with length 6 that has:

- The letter (و, WAW) as the second letter;
- The letter (ا, ALIF) as the third letter;
- The letter (ي, YAA) as the fifth letter;

This condition represents only the pattern "فواعيل" (FAWAAIL). Next step, we remove these letters and extract the root "حسب" (Compute) (Fig 5).

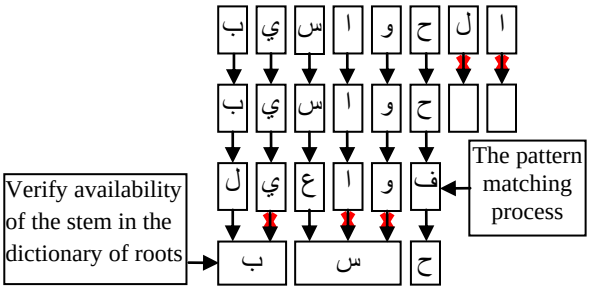

Fig 5: Example of the method PS+M.

The order of these rules and conditions used for comparison are very important factors to ensure a correct comparison.

2.5.1.4 The method SP+M (Prefix or Suffix with Pattern/Model):

SP+M based on the same principle of the method PS+M, but by removing the suffixes firstly and the suffixes secondly.

2.5.1.5 The method HY (Hybrid):

Since each method of stemming has its limits, it is natural to consider combining them to benefit from the advantages of each of them.

The proposed hybrid method integrates three different stemming techniques, including: affix-removal, morphological analysis and dictionaries. The global structure of the program of this method is given as follows:

```
Program
  While there is a prefix do
    Check the existence of the word in the dictionary //The
    technical dictionaries
    If it exists then
      Add to index
      Exit program
    Else
      While there are suffixes do
        While there are models do
          Compare the word with the model //The
          morphological analysis technique
          If there is a model then
            Extract the root
            Check the existence of the word in the dictionary
            //The technical dictionaries
            If it exist then
                Add to index
                Exit program
            End If
          End If
        End While
        Remove the suffix //The affix-removal dictionaries
      End While
    End Else
    Remove the prefix //The affix-removal dictionaries
  End while
```

Wrong word
End Program

Through the combination of stemming techniques, we have improved the index quality of the corpus (documents and queries), and consequently we have a good performance of the Arabic IR.

3. EXPERIMENTATION AND EVALUATION

Stemming is necessary for the performance of IR; it allows combining terms which have similar meaning, with small differences in the morphological form having a single index, and therefore it improves the quality of retrieval.

The aim of our experiments is to evaluate different methods of stemming performance in Arabic information retrieval. A series of experiments was conducted to show the effect of each method of stemming in retrieval performance.

We first compare the two methods of stemming (PS-M and SP-M) that we have proposed. Bellow Figure 6 represents a comparison between these two methods based on their recall-precision curves. The results show that the method of stemming SP-M is frequently more efficient than the PS-M on points of recall; curve SP-M representing the precision of search based on points of recall is frequently above the PS-M curve.

Also, these results show that PS-M stemming for Arabic words is not the best method for Arabic IR, while the method SP-M can better determine the semantic core of a word and therefore it increases the performance of IR.

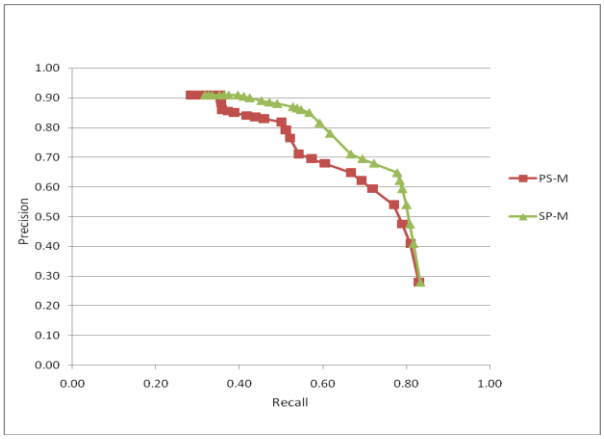

Fig 6: Curves recall-precision of the two methods of stemming PS-M and SP-M.

Then, we compare the two methods of stemming (PS+M and SP+M). Because stemming based on these two methods proceeds differently, we introduce a new factor which is the pattern (AOUZANE). This pattern produces a set of candidate stems using the dictionary of roots to choose the better one.

To compare the PS+M method with the SP+M method, we plot the recall-precision curve. Figure 7 represents a comparison between these two methods of stemming.

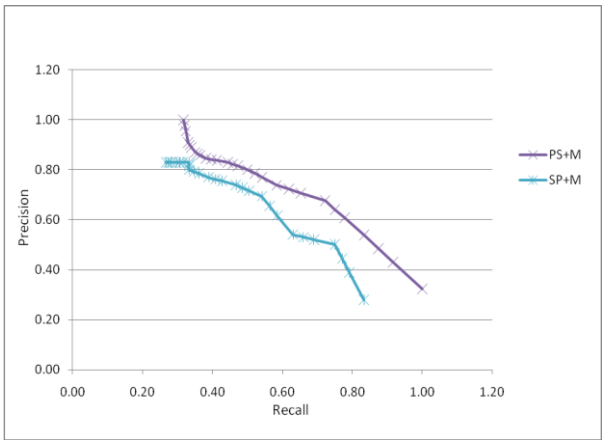

Fig 7: Curves recall-precision of the two methods of stemming PS+M and SP+M.

Unlike the method of stemming without pattern, the results show that the method of stemming PS+M is more efficient than the SP+M on all points of recall; curve PS+M representing the precision of retrieval based on the points of recall is always above the curve SP+M

Also, these results show that the method PS+M can better determine the semantic core of a word, and therefore it increases the performance of IR.

Figure 8 represents a comparison between the stemming methods without pattern and those with patterns according to their curves recall-precision.

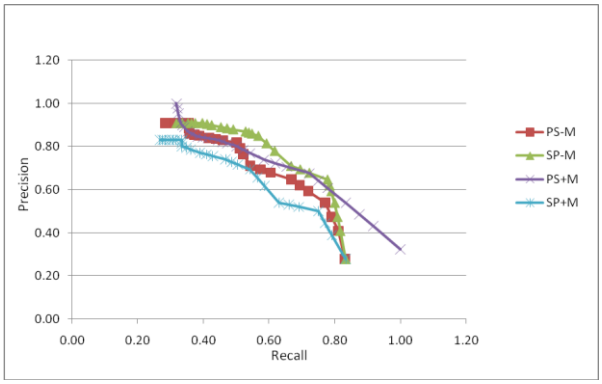

Fig 8: Curves recall-precision of stemming methods PS-M, SP-M, PS+M and SP+M.

In our experiments series, the results show that the tow methods of stemming SP-M and PS+M are more effectives than other methods (PS-M and SP+M). One can observe this behavior in Figure 9; the curve stemming SP-M representing the precision of search based on points of recall is often above the other curves. In all the test queries, we obtain 57% average accuracy with the method of stemming SP-M against 48%, 51% and 54% for methods SP+M, PS-M and PS+M respectively.

However, this Figure 8 shows that the method of stemming PS+M gets the best scores when the recall is less than 35%.

For this reason we propose a new hybrid method of stemming (HY) which combines all the methods mentioned above and improves the overall performance of the process of stemming.

Figure 9 bellow illustrates a comparison between the five stemming methods according to their recall-precision curves.

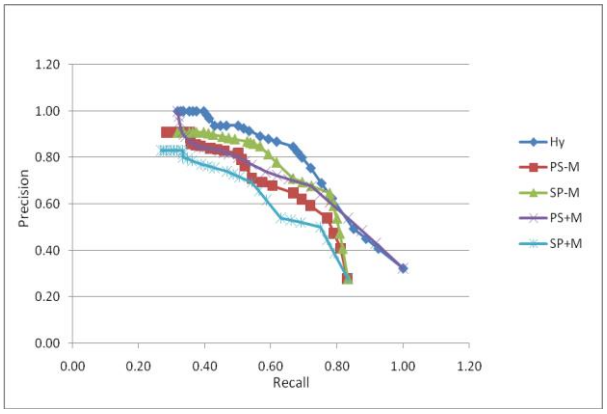

Fig 9: Curves recall-precision of the five methods of stemming.

These results show that the method of stemming HY is more effective than the other methods (SP-M, PS-M, PS+M and SP+M). One can observe this behavior in Figure 9; stemming curve HY representing accuracy of retrieval based on recall points is often above the other curves. We obtain 58% average accuracy with the HY method of stemming and 57% with the method SP-M against 48%, 51% and 54% for methods M+SP, SP-M and PS+M respectively.

Also, these results show that the HY method of stemming is the best approach as it allows more successfully to group semantically similar words. Yet the other methods cannot determine the best semantics core of a word.

The HY method of stemming does not make a blind truncation; it applies different decompositions processes onto the original word. In the presence of multiple affixes in a word, it allows to successfully choose the affix (prefix or suffix) to eliminate, and to correctly determine the appropriate pattern. And as a result, it allows to produce a set of candidate stems using the dictionary of roots, and to choose the better one stem.

The method of stemming HY is not perfect, and cannot identify the correct stem for some ambiguous words; it is in this aspect that our method must be improved.

4. CONCLUSION

Arabic is one among the most widely used languages in the world, but relatively speaking; only few studies are made on Arabic information retrieval and Arabic texts classification.

The main objective of this work is to implement an IR system on Arabic textual documents, to test some methods of stemming and evaluate these methods. Several methods are largely invested in a number of text processing and information retrieval. The main problem of each proposed method is how to identify the best index terms for reasonable performance.

In this context, we propose a methodology for improving the research performance of Arabic textual documents. We apply five different methods of stemming, and compare their results. So we came to conclude which method gives better performance in information retrieval. Indeed, among these five methods of stemming, we propose a new hybrid method of stemming (HY).With this method we try to determine the core of a word by the integration of three different techniques (deletion of affix, dictionaries, and morphological analysis) to improve the overall performance of the process of stemming. This new method has better search performance than the other methods because it can better determine the stem of a word. Thus, we have shown the effectiveness of removing suffixes before prefixes during the stemming process of Arabic texts. Yet the new method can also cause errors due to ambiguity. The presence of a few cases of ambiguity in this method does not pose problems for Arabic information retrieval because these same words in the texts are stemmed the same way, and therefore their identified stems are identical to those obtained for words in queries.

Other cases of errors occurring sometimes, when terms are not semantically similar, are grouped into an equivalence class. It is in this aspect that our method must be improved.

5. REFERENCES

[1] Kadri, Y. & Nie, J. (2006). "Effective Stemming for Arabic Information Retrieval" in proceedings of the Challenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni.

[2] Al-Kharashi, I. and Evens, M. W. Comparing words, stems, and roots as index terms in an Arabic information retrieval system. JASIS, 45 (8), pp. 548-560, 1994.

[3] Kenneth R. Beesley. 1998. Arabic Morphological Analysis on the Internet. To appear in the Proceedings of the International Conference and Exhibition on Multi-lingual Computing (Arabic and English), ICEMCO-98.

[4] Attia, Mohamed, A.: 2000 A large-scale computational processor of the Arabic morphology, A Master's Thesis, Cairo University, (Egypt) (2000).

[5] Mohamadi, T.S. Mokhnache: 2002, Design and development of Arabic speech synthesis, WSEAS 2002, Greece, Sept. 25-28, (2002).

[6] http://www.internetworldstats.com/stats.htm.

[7] Khoja S. and Garside S. (1999). 'Stemming Arabic Text'. Computing Department, Lancaster University, Lancaster, U.K.

[8] Larkey L. S. and Connell M. E. (2001). 'Arabic information retrieval at UMass in TREC-10'. TREC-10 conference, Gaithersburg, Maryland 2001.

[9] Darwish, K. and Oard, D.W. CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval. In TREC 2002. Gaithersburg: NIST, pp 703-710, 2002.

[10] Chen, A., and Gey, F. Building an Arabic stemmer for information retrieval. In TREC 2002. Gaithersburg: NIST, pp 631-639, 2002.

[11] Wightwick, J. and Gaafar, M. Arabic verbs and essentials of grammar. Chicago: Passport Books, 1998.

[12] Larkey L.S, L. Ballesteros, and M.E. Connell, "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis," Tampere, Finland: ACM, 2002, pp. 275-282.

[13] P. Schauble, Multimedia Information Retrieval: content-based Information Retrieval from Large Text and Audio Databases, Kluwer Academic Publishers, 1997

[14] Pirkola, A. Morphological typology of languages for IR. Journal of Documentation, 57 (3), pp. 330-348, 2001.

[15] Popovic, M. and Willett, P. The effectiveness of stemming For natural-language access to Slovene textual data. JASIS, 43 (5), pp. 384-390, 1992.

[16] Ntais, G. Development of a stemmer for the greek language. Master's thesis, Stockholm University, 2006.

[17] Sankupellay, M. "Malay-Language Stemmer," Sunway Academic Journal, vol. 3, pp. 147–153, 2006.

[18] Al-Sughaiyer, I. A. and Al-Kharashi, I. A. (2004) "Arabic morphological analysis techniques: A comprehensive survey", Journal of the American Society for Information Science and Technology, 55(3):189–213.