LREC'2012 Workshop: LRE-Rel

Language Resources and Evaluation for Religious Texts

LRE-Rel Workshop Programme

Tuesday 22 May 2012

09:00 – 10:30 – Session 1 Papers

09:00 Eric Atwell, Claire Brierley, and Majdi Sawalha (Workshop Chairs)
Introduction to Language Resources and Evaluation for Religious Texts

09.10 Harry Erwin and Michael Oakes
Correspondence Analysis of the New Testament

09.30 Mohammad Hossein Elahimanesh, Behrouz Minaei-Bidgoli and Hossein Malekinezhad
Automatic classification of Islamic Jurisprudence Categories

09.50 Nathan Ellis Rasmussen and Deryle Lonsdale
Lexical Correspondences Between the Masoretic Text and the Septuagint

10.10 Hossein Juzi, Ahmed Rabiei Zadeh, Ehsan Baraty and Behrouz Minaei-Bidgoli
A new framework for detecting similar texts in Islamic Hadith Corpora

10:30 – 11:20 Coffee break and Session 2 Posters

Majid Asgari Bidhendi, Behrouz Minaei-Bidgoli and Hosein Jouzi
Extracting person names from ancient Islamic Arabic texts

Assem Chelli, Amar Balla and Taha Zerrouki
Advanced Search in Quran: Classification and Proposition of All Possible Features

Akbar Dastani, Behrouz Minaei-Bidgoli, Mohammad Reza Vafaei and Hossein Juzi
An Introduction to Noor Diacritized Corpus

Karlheinz Mörth, Claudia Resch, Thierry Declerck and Ulrike Czeitschner
Linguistic and Semantic Annotation in Religious Memento Mori Literature

Aida Mustapha, Zulkifli Mohd. Yusoff and Raja Jamilah Raja Yusof
The Qur'an Corpus for Juzuk Amma

Mohsen Shahmohammadi, Toktam Alizadeh, Mohammad Habibzadeh Bijani and Behrouz Minaei
A framework for detecting Holy Quran inside Arabic and Persian texts

Gurpreet Singh
Letter-to-Sound Rules for Gurmukhi Panjabi (Pa): First step towards Text-to-Speech for Gurmukhi

Sanja Stajner and Ruslan Mitkov
Style of Religious Texts in 20th Century

Daniel Stein
Multi-Word Expressions in the Spanish Bhagavad Gita, Extracted with Local Grammars Based on Semantic Classes

Nagwa Younis
Through Lexicographers' Eyes: Does Morphology Count in Making Qur'anic Bilingual Dictionaries?

Taha Zerrouki, Ammar Balla
Reusability of Quranic document using XML

11:20 – 13:00 – Session 3 Papers

11.20 Halim Sayoud
Authorship Classification of two Old Arabic Religious Books Based on a Hierarchical Clustering

11.40 Liviu P. Dinu, Ion Resceanu, Anca Dinu and Alina Resceanu
Some issues on the authorship identification in the Apostles' Epistles

12.00 John Lee, Simon S. M. Wong, Pui Ki Tang and Jonathan Webster
A Greek-Chinese Interlinear of the New Testament Gospels

12.20 Soyara Zaidi, Ahmed Abdelali, Fatiha Sadat and Mohamed-Tayeb Laskri
Hybrid Approach for Extracting Collocations from Arabic Quran Texts

12.40 Eric Atwell, Claire Brierley, and Majdi Sawalha (Workshop Chairs)
Plenary Discussion

13:00 End of Workshop

Editors

Eric Atwell:	University of Leeds, UK
Claire Brierley:	University of Leeds, UK
Majdi Sawalha:	University of Jordan, Jordan

Workshop Organizers

AbdulMalik Al-Salman:	King Saud University, Saudi Arabia
Eric Atwell:	University of Leeds, UK
Claire Brierley:	University of Leeds, UK
Azzeddine Mazroui:	Mohammed First University, Morocco
Majdi Sawalha:	University of Jordan, Jordan
Abdul-Baquee Muhammad Sharaf:	University of Leeds, UK
Bayan Abu Shawar:	Arab Open University, Jordan

Workshop Programme Committee

Nawal Alhelwal:	Arabic Department, Princess Nora bint Abdulrahman University, Saudi Arabia
Qasem Al-Radaideh:	Computer Information Systems, Yarmouk University, Jordan
AbdulMalik Al-Salman:	Computer and Information Sciences, King Saud University, Saudi Arabia
Eric Atwell:	School of Computing, University of Leeds, UK
Amna Basharat:	Foundation for Advancement of Science and Technology, FAST-NU, Pakistan
James Dickins:	Arabic and Middle Eastern Studies, University of Leeds, UK
Kais Dukes:	School of Computing, University of Leeds, UK
Mahmoud El-Haj:	Computer Science and Electronic Engineering, University of Essex, UK
Nizar Habash:	Center for Computational Learning Systems, Columbia University, USA
Salwa Hamada:	Electronics Research Institute, Egypt
Bassam Hasan Hammo:	Information Systems, King Saud University, Saudi Arabia
Dag Haug:	Philosophy, Classics, History of Art and Ideas, University of Oslo, Norway
Moshe Koppel:	Department of Computer Science, Bar-Ilan University, Israel
Rohana Mahmud:	Computer Science & Information Technology, University of Malaya, Malaysia
Azzeddine Mazroui:	Mathematics and Computer Science, Mohammed 1st University, Morocco
Tony McEnery:	English Language and Linguistics, University of Lancaster, UK
Aida Mustapha:	Computer Science and Information Technology, University of Putra, Malaysia
Mohamadou Nassourou:	Computer Philology & Modern German Literature, Uni of Würzburg, Germany
Nils Reiter:	Department of Computational Linguistics, Heidelberg University, Germany
Abdul-Baquee M. Sharaf:	School of Computing, University of Leeds, UK
Bayan Abu Shawar:	Information Technology and Computing, Arab Open University, Jordan
Andrew Wilson:	Linguistics and English Language, University of Lancaster, UK
Nagwa Younis:	English Department, Ain Shams University, Egypt
Wajdi Zaghouani:	Linguistic Data Consortium, University of Pennsylvania, USA

Table of contents

Author Index

Introduction to
Language Resources and Evaluation for Religious Texts

Eric Atwell, Claire Brierley, and Majdi Sawalha (Workshop Chairs)

Welcome to the first LRE-Rel Workshop on Language Resources and Evaluation for Religious Texts, part of the LREC'2012 Language Resources and Evaluation Conference in Istanbul, Turkey. The focus of this workshop is the application of computer-supported and Text Analytics techniques to religious texts ranging from: the faith-defining religious canon; authoritative interpretations and commentary; sermons; liturgy; prayers; poetry; and lyrics. We see this as an inclusive and cross-disciplinary topic, and the workshop aims to bring together researchers with a generic interest in religious texts to raise awareness of different perspectives and practices, and to identify some common themes.

We therefore welcomed submissions on a range of topics, including but not limited to:

- analysis of ceremonial, liturgical, and ritual speech; recitation styles; speech decorum; discourse analysis for religious texts;
- formulaic language and multi-word expressions in religious texts;
- suitability of modal and other logic types for knowledge representation and inference in religious texts;
- issues in, and evaluation of, machine translation in religious texts;
- text-mining, stylometry, and authorship attribution for religious texts;
- corpus query languages and tools for exploring religious corpora;
- dictionaries, thesaurai, Wordnet, and ontologies for religious texts;
- measuring semantic relatedness between multiple religious texts;
- (new) corpora and rich and novel annotation schemes for religious texts;
- annotation and analysis of religious metaphor;
- genre analysis for religious texts;
- application in other disciplines (e.g. theology, classics, philosophy, literature) of computer-supported methods for analysing religious texts.

Our own research has focussed on the Quran (e.g. see Proceedings of the main Conference, LREC'2012); but we were pleased to receive papers dealing with a range of other religious texts, including Muslim, Christian, Jewish, Hindu, and Sikh holy books, as well as religious writings from the 17th and 20th centuries. Many of the papers present an analysis technique applied to a specific religious text, which could also be relevant to analysis of other texts; these include text classification, detecting similarities and correspondences between texts, authorship attribution, extracting collocations or multi-word expressions, stylistic analysis, Named Entity recognition, advanced search capabilities for religious texts, developing translations and dictionaries.

This LRE-Rel Workshop demonstrates that religious texts are interesting and challenging for Language Resources and Evaluation researchers. It also shows LRE researchers a way to reach beyond our research community to the billions of readers of these holy books; LRE research can have a major impact on society, helping the general public to access and understand religious texts.

Extracting person names from ancient Islamic Arabic texts

Majid Asgari Bidhendi, Behrouz Minaei-Bidgoli, Hosein Jouzi

School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
Computer Research Center of Islamic Sciences, Qom, Iran
majid_asgari@comp.iust.ac.ir, b_minaei@iust.ac.ir, hjuzi@noornet.net

Abstract

Recognizing and extracting name entities like person names, location names, date and time from an electronic text is very useful for text mining tasks. Named entity recognition is a vital requirement in resolving problems in modern fields like question answering, abstracting systems, information retrieval, information extraction, machine translation, video interpreting and semantic web searching. In recent years many researches in named entity recognition task have been lead to very good results in English and other European languages; whereas the results are not convincing in other languages like Arabic, Persian and many of South Asian languages. One of the most necessary and problematic subtasks of named entity recognition is person name extracting. In this article we have introduced a system for person name extraction in Arabic religious texts using proposed "Proper Name candidate injection" concept in a conditional random fields model. Also we have created a corpus from ancient Arabic religious texts. Experiments have shown that very hight efficient results have been obtained based on this approach.

Keywords: Arabic Named entity recognition, information extraction, conditional random fields

1. Introduction

Named entity identification that also known as Named entity recognition and name entity extraction, is a subtask of information extraction and that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations (companies, organizations and etc.), locations (cities, countries, rivers and etc.), time and dates, quantities, etc. As we have shown in next section, named entity extraction task and especially person name extracting haven't been lead to convincing results in Arabic language. Moreover, most of works which done for NER in Arabic language, have been focused on modern newspaper data. As we will show, there are very significant differences between newspaper data and ancient religious texts in Arabic language. In this paper we have focused specially on NER task for three main type of Islamic texts: historic, Hadith[1] and jurisprudential books. Person name extracting is very useful for Islamic religious sciences. Especially, for historic and Hadith books, finding the relationships between person names is a very valuable task. In Hadith books, people cited quotations from main religious leaders of Islam. These valuable data can help us to verify correctness of citations based on known truthful and untruthful relaters. Also we must point out that NER task is very useful subtask in text processing (and also in religious text processing) which can help other subtasks of natural language processing(Benajiba et al., 2004).

The rest of this paper is structured as follows. In section 2., we investigate the other efforts have been made to solve named entity task in Arabic language. In section 3., we emphasize on making a proper corpora for Arabic religious texts, NoorCorp, and NoorGazet, a gazetteer which contains religious person names. Section 4., explains in details Noor ANER system based on "proper name candidate injection" concept in conditional random fields model. Finally, in section 5.. we present the experiments we have carried out with our system, whereas in last section we draw our conclusions and future works.

2. Related works

Before 2008, the most successful documented and comparable efforts had been made in named entity recognition task was ANERsys (Benajiba et al., 2007) (based on maximum entropy), ANERsys 2 (Benajiba and Rosso, 2007) (based on conditional random fields) and proprietary Siraj system. In 2008, Benjiba and Rosso, proposed an another system based on combining results from four conditional random fields models (Benajiba and Rosso, 2008). Afterward Aljomaily et al. introduced an online system based on pattern recognition in (Al-Jumaily et al., 2011). Those patterns were built up by processing and integrating different gazetteers, from DBPedia (http://dbpedia.org/About, 2009) to GATE (A general architecture for text engineering, 2009) and ANERGazet (http://users.dsic.upv.es/grupos/nle/?file=kop4.php). All of these efforts have presented their results on ANERCorp that has been made by (Benajiba et al., 2007). The best results in extracting person named was obtained in (Al-Jumaily et al., 2011). They recorded F-measure equals to 76.28 for person names in their results. Furthermore some efforts have been made on NER task in specific domains. For example in (Fehri et al., 2011), F-measure equals to 94 has been obtained for sport news. In (Elsebai and Meziane, 2011), authors presented a method based on using a keyword set instead complicated syntactic, statistical or machine learning approaches.

[1] The term Hadith is used to denote a saying or an act or tacit approval or criticism ascribed either validly or invalidly to the Islamic prophet Muhammad or other Islamic leaders. Hadith are regarded by traditional Islamic schools of jurisprudence as important tools for understanding the Quran and in matters of jurisprudence.

3. Preparing NoorCorp and NoorGazet

As reported in Conference on Computational Natural Language Learning in 2003 (Tjong Kim Sang and De Meulder, 2003), a corpora for named entity task must contains words with theirs NER tags. Same classes those was defined in Message Understanding Conference, contains organizations, locations and person names. Other types of named entities are tagged with MISC. Therefore, each word must be tagged with one of these tags:

- B-PERS: Beginning of a person name.
- I-PERS: Inside or ending of a person name.
- B-LOC: Beginning of a location name.
- I-LOC: Inside or ending of a location name.
- B-ORG: Beginning of a organization name.
- I-ORG: Inside or ending of a organization name.
- B-MISC: Beginning of a name which is not a person, location or organization name.
- I-MISC: Inside or ending of a name which is not a person, location or organization name.
- O: Other words.

In CoNLL, the decision has been made to keep a same format for training and test data for all languages. The format consists of two columns: the first one for words and second one for tags. Figure 1 shows two examples of standard corpora for NER task. In left side, a piece of tagged English text is shown. Often a TAB or SPACE character is used between word and its tag. Each word is written with its tag or tags in one separate line. All punctuation marks are written in separate line just like the other words. To make a proper corpora for NER task in Arabic religious texts, 3 corpora have been prepared from tagged text in Computer Research Center of Islamic Sciences which is located in Qom, Iran. These 3 corpora have been made from 3 books:

```
Arsenal    B-ORG    |    0       و
captain    0        |    0       جلس
Robin      B-PER    |    B-PER   سليمان
van        I-PER    |    0       قليلا
Persie     I-PER    |    0       ثم
has        0        |    0       نهض
spoken     0        |    0       ف
of         0        |    0       خرج
his        0        |    0       إلى
love       0        |    B-PE    حسن
for        0        |    I-PER   بن
London     B-LOC    |    I-PER   على
,          0        |    0       و
      ...           |    0       هو
                    |    0       قاعد
                    |    0       في
                    |    B-LOC   المسجد
                    |    B-LOC   الكوفة
```

Figure 1: Standard English and Arabic corpora for NER task

- A historical book, "Vaghat-a Seffeyn" written by "Nasr bin Mozahim Manghari" in 828 A.D.
- A traditional Hadith book, "El-Irshad fi marefati hojajellah alal-ibad" written by "Mohammad bin Mohammad Mofid" in 846 A.D.
- A jurisprudential book, "Sharaye el-Islam fi masaele harame val-halal" written by "Jafar bin Hasan" in 1292 A.D.

These corpora are compared based on number of words (after tokenizing), ratio of person, location and organization names in table 1. Also those are compared with ANERCorp (which contains newspaper data) in that table.
Table 1 shows these 4 corpora which belongs to modern and ancient Arabic text, are very different in their structures. In addition to differences between modern and ancient Arabic texts, ratio of using person and location names are different. In newspaper data, this ratio is high, but in compare to historical data, ratio of using names are lower than historical text. In traditional Hadith book, ratio of person names are higher, but names of locations are lower than mentioned texts. In jurisprudential texts, ratio of proper names are very low. In that corpora, the portion of proper names is lesser than 1 percent (totally 233 proper names in 48582 words). Ratio of each type of proper names are shown in table 2.

Corpus	Person	Location	Organization	Misc
Seffeyn	27.98%	43.52%	28.5%	0%
El-Irshad	80.66%	6.03%	13.03%	0%
Sharaye	30.19%	69.81%	0%	0%
ANERcorp	38.98%	30.42%	20.58%	10.01%

Table 2: Ratio of each types of proper names in NoorCorp and ANERCorp

Gazetteers are many important resources to improve the results of NER task. To make a perfect gazetteer, about 88000 proper names were gathered from "Jamiál-AHadith" software which has been produced by Computer Research Center of Islamic Sciences. Then we tokenized these names to their elements. For example "Hasan bin Ali bin Abdellah bin el-Moghayrah" was tokenized to 6 unrepeated elements: "Hasan", "bin", "Ali", "Abd", "Allah" and "Moghayrah". These elements were produced for all proper names and added to a database with their frequencies. Finally a database with 18238 names were produced.

4. Noor ANER System

Noor ANER is a system based on conditional random fields which analyzes input text and extracts proper names after three types of preprocessing. We describe Noor ANER and its structure in this section.

4.1. Conditional Random Fields

Conditional random fields is a statistical modeling method which often is used in pattern recognition. Precisely, CRF is a discriminative undirected probabilistic graphical model.

4.1.1. Log-Linear Models

Let x as an example and y as a possible tag for that. A log-linear model supposes

$$p(y|x;w) = \frac{e^{\sum_j w_j F_j(x,y)}}{Z(x,w)} \qquad (1)$$

Corpus	Number of words	Person	Location	Organization	Misc	Subject
Seffeyn	235842	6.47%	10.06%	6.59%	0%	History
El-Irshad	134316	14.31%	1.07%	2.36%	0%	Hadith
Sharaye	48582	0.48%	1.11%	0%	0%	jurisprudence
ANERcorp	150285	4.28%	3.34%	2.26%	1.10%	Newspaper

Table 1: NoorCorp and its containing books.

that Z is named as "partition function" and it equals with

$$Z(x,w) = \sum_{y'} e^{\sum_j w_j F_j(x,y')} \tag{2}$$

Therefore, having input x, predicted tag from model will be

$$\hat{y} = argmax_y p(y|x;w) = argmax_y \sum_j w_j F_j(x,y) \tag{3}$$

each of $F_j(x,y)$ are feature functions.
CRF model are a specific type of Log-Linear models. CRF in this article, refers to Linear-chain CRF.

4.1.2. Induction and training in CRF models

Training of CRF model means finding wight vector w such that make best possible prediction for each training example $\bar{x}$:

$$\bar{y}^* = argmax_{\bar{y}} p(\bar{y}|\bar{x};w) \tag{4}$$

However, before describing training phase, we must consider two main problems exists in induction phase: First, how can we compute 4 equation for each $\bar{x}$ and each set of weights w efficiently. This computation is exponential due to number of different sequences for tags $\bar{y}$. second, having $\bar{x}$ and $\bar{y}$ we must evaluate these values:

$$p(\bar{y}|\bar{x};w) = \frac{1}{Z(\bar{x},w)} e^{\sum_j w_j F_j(\bar{x},\bar{y})} \tag{5}$$

problem in here is denominator, because that needs all of sequences $\bar{y}$:

$$Z(\bar{x},w) = \sum_{\bar{y}'} e^{\sum_j w_j F_j(\bar{x},\bar{y}')} \tag{6}$$

for both of these problems, we needs efficient innovative methods, which without moment processing on each $\bar{y}$ in 6, processes all of them efficiently. The assumption that each feature function in this CRF models are dependent to two adjacent tags, aim us to resolve this problems. You can refer to (Elkan, 2008), (Lafferty et al., 2001) or (Sutton and McCallum, 2007) for more information.
When we have a set of training examples, we suppose our goal is finding parameters w_j so that conditional probability of occurring those training examples would be maximum. For this propose, we can use ascending gradient method. Therefore we need to compute conditional likelihood for a training example for each w_j. maximizing p is same as maximizing $ln\, p$:

$$\begin{aligned}\frac{\partial}{\partial_{w_j}} ln\, p(y|x;w) &= F_j(x,y) - \frac{\partial}{\partial_{w_j}} logZ(x,w) \\ &= F_j(x,y) - E_{y' \sim p(y'|x;w)}[F_j(x,y')].\end{aligned} \tag{7}$$

In other words, partially derivation to ith weight is value of ith feature function for true tag y minus average value of feature function for all of possible tags y'. Note that this derivation allows real value for each feature function, not only zero and one values. When we have all of training examples T, gradient ascending of condition likelihood, will be the sum of ascending for each training examples. Absolute maximum of all of these ascending are equal to zero, Therefore:

$$\sum_{\langle x,y \rangle \in T} F_j(x,y) = \sum_{\langle x,. \rangle \in T} E_{y \sim p(y'|x;w)}[F_j(x,y)] \tag{8}$$

This equation is correct for all of training examples not for each of them. Left side of above equation is total value of feature function j on all of training sets. Right side is total value of feature function j which is predicted by model. Finally, when we maximize conditional likelihood with online ascending method, adjustment of weight w_j would be calculated with this formula:

$$w_j := w_j + \alpha(F_j(x,y) - E_{y' \sim p(y'|x;w)}[F_j(x,y')]) \tag{9}$$

4.2. Preprocessing methods

In training, testing and prediction phases we are using some preprocessing methods. In this section we describe about these methods.

4.2.1. Tokenizing

One of must useful preprocessing on text mining tasks, are tokenization. Tokenization is the process of breaking text up into words, phrases, symbols, or other **meaningful** elements called **tokens**. For example the word "Sayaktobounaha" in Arabic language, (which means "And they will write that") will be tokenized into "va+sa+ya+ktob+ooua+ha".
We have used AMIRA 2.1 software for tokenization process. We will describe more about that software in section 4.2.3..

4.2.2. Transliteration

Another useful preprocessing method which often is last preprocess, is transliteration. Transliteration is replacing characters of first language with character of a destination language. Often second language is English. In this process, each character is mapped to one and just one character in destination language. In Noor ANER system we used Buckwalter transliteration. Figure 2 shows mentioned transliteration for Arabic and Persian languages (Habash et al., 2007). Figure 3 shows some examples for this transliteration. Second column from right, is real data in Arabic language and first column is transliterated data. Many general proposed language processing tools accept their inputs

'	ء	x	خ	-	ـ	K	ـٍ
|	آ	d	د	f	ف	a	ـَ
>	أ	*	ذ	q	ق	u	ـُ
&	ؤ	r	ر	k	ك	i	ـِ
<	إ	z	ز	l	ل	~	ـّ
}	ئ	s	س	m	م	o	ـْ
A	ا	$	ش	n	ن	P	پ
b	ب	S	ص	h	ه	J	چ
p	ة	D	ض	w	و	V	ژ
t	ت	T	ط	Y	ى	G	گ
v	ث	Z	ظ	y	ي		
j	ج	E	ع	F	ـً		
H	ح	g	غ	N	ـٌ		

Figure 2: Buckwalter transliteration for Arabic language

in first column. For this reason the transliterated data is placed there.

```
w          و     CC          CC         O
jls      جلس     VBD_MS3     VP         O
slymAn سليمان    NNP         NP    B-PERS
qlylA    قليلا   NN        ADJP         O
vm        ثم     CC          CC         O
nhD      نهض     VBD_MS3     VP         O
f          ف     RP          RP         O
xrj      خرج     VBD_MS3     VP         O
<ly      إلى     IN          PP         O
AlHsn   الحسن    DET_NNP     NP    B-PERS
bn        بن     NNP         NP    I-PERS
Ely      علي     NNP         NP    I-PERS
w          و     CC          CC         O
hw        هو     PRP_MS3     NP         O
qAEd    قاعد     NN          NP         O
fy        في     IN          PP         O
Almsjd المسجد    DET_NN      NP     B-LOC
Alkwfh الكوفه    DET_JJ      NP     I-LOC
```

Figure 3: Corpora after transliteration and adding POS and BPC tags

4.2.3. AMIRA software and part of speech tagging

AMIRA software has been developed by Mona Diab in Colombia University for standard Arabic language. AMIRA is a replacement for ASVMTools. This software contains a Tokenizer (TOK), a part of speech tagger (POS) and a base phrase chunker (BPC). The reports which were published in (Diab, 2009) shows this toolkit is very fast and reliable. Also user can adjust many different parameters in this software. AMIRA has been used in many papers about natural language processing in Arabic language. We have used this software toolkit in preprocessing phases of Noor ANER system.

4.2.4. Corpus preparation and training of CRF model

In Noor ANER system we have used FlexCRF, a general proposed implementation of conditional random fields. That software accepts the input in the following structure: The input must has three columns:

- First column, contains transliterated data. In our problem sequences for tagging process are sentences. Each sentence must ends with a period character. after each sentence, one line leaves blank.
- Second column consists feature functions. Structure of these feature functions has been described in documentation files of this software[2]. We are free to use any valid feature function sets for this column. But we must meet limitations of conditional random fields model. Therefore each feature function must depends on current word or predicate, up to two previous words or predicates and up to two next words or predicates. Our system uses these feature function templates:
  - One word.
  - Two consecutive words.
  - One predicate.
  - Two consecutive predicates.
  - Three consecutive predicates.
  - One word and one predicate.
  - Two predicates and one word.
  - Two words and one predicate.

  Predicates in Noor ANER are POS tags of each words. These POS tags are assigned by AMIRA software.
- Third column is NER tag for training and testing phases of CRF model.

As you can see in figure 3, we have different information for each word:

- Transliterated words (generated from original text).
- Original words (Typed by typists).
- POS tags (generated by AMIRA software from original text).
- BPC tags (generated by AMIRA software from original text).
- NER tags (verified by linguists)

The last column is needed for training and testing phases not in prediction phase.

4.3. Proper Name Candidate Injection

We described in previous section that predicates used in training of CRF model are POS tags. But indeed, predicates **are not exactly** POS tags. We have adjusted POS tags to improve the results in NER task. We enrich POS tags which are generated by AMIRA software from original input text:

1. If current word, is existed in our gazetteer, "NAME_" phrase is added to beginning of it POS tag. We named this word a "Proper Name Candidate".
2. If we encountered to two or more consecutive proper name candidates, we replace the POS tag with "NAME2" tag.

[2] Refer to this address for more information: http://flexcrfs.sourceforge.net/documents.html

In this approach, if total number of POS tags are n, the size of predicates will be $2n + 1$.

Why we expect better results with this approach? Importance of second item seems obvious. Many person names in Arabic languages also have adjective roles. But in major cases, when two or more of these word placed consecutively, we can tagged those as proper names, with very high probability. Especially, existence of relational words like "Bin" between them, raises this possibility. This probability was 94 percent in our experiments and based on this fact, we replace the POS tag to a new constant NAME2 tag here. First rule is very useful too. In fact we are producing a predicate that consists of POS information and a proposal to be a proper name. However, CRF model is deciding to tag this word as proper name or not. Using this approach, we generate extra predicates which have more probability to be proper names. But yet this is the CRF model which decides how to use this new informations.

With these descriptions, we expect to gain a reliable approach. Because when the POS tagger, AMIRA or any other software, tagged one word wrongly, the CRF models just ignores this tag sequence in training phase, because it don't find any more sentences with this wrong POS tag sequence. (Ignorance don't mean a complete ignorance here. CRF saves all of feature functions. but the possibility of using this new wrong tag sequence is very very low for a huge corpora) **Our experiences proved this claim**.

4.4. Structure of Noor ANER System

As we mentioned above, our tagged texts are converted to an standard transliterated NER corpora by some preprocessing tools. Then, we produced text with POS and NER tags using POS tagger software. Then another software generates predicates which enriched with proper name candidates. Generated resource after these processes is delivered to CRF trainer. Figure 4 shows this structure. In prediction phase, we have same process, but no NER tag is available in the resources.

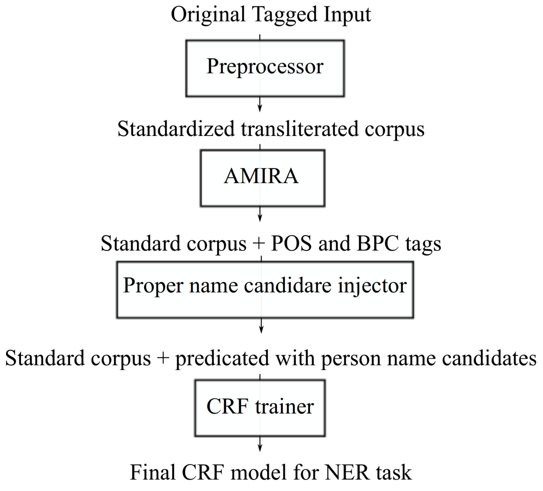

Original Tagged Input
Preprocessor
Standardized transliterated corpus
AMIRA
Standard corpus + POS and BPC tags
Proper name candidare injector
Standard corpus + predicated with person name candidates
CRF trainer
Final CRF model for NER task

Figure 4: Structure of Noor ANER System

Corpus	Topic	Precession	Recall	F-measure
Seffeyn	History	99.93%	99.93%	99.93
Al-Irshad	Hadith	95.62%	92.16%	93.86
Sharaye	Jurisprudence	100.00%	60.87%	75.68

Table 3: Evaluation of Noor ANER system on NoorCorp

5. Evaluation

We introduced 3 new corpora in section which are produced for Arabic NER task. Those corpora contains 3 different topics: history, Hadith and jurisprudence. Since Noor ANER has focused on person names, the results are shown just for person names. As table 3 shows, precession and recall metrics are very high for historical and traditional Hadith data. One of most reasons to obtain this high accuracy is existence of full names (that contains first name. father's name, ancestors name, nickname and even last name) in these topics. And full names consists their parts which are connected with some frequent relational words like "Bin". Therefore CRF model has a very strong pattern to extract many of person names.

Proper names in jurisprudence data are rare, thus extracting person names in this case is very very hard and not reliable. The results shows this fact.

6. Conclusion and future works

Results showed that Noor ANER act with very good performance on religious texts. The experiments declared we have very high F-measure for historical and Hadith data. Also we have produced 3 corpora based on three religious books in Arabic languages.

it is important to point out that we have used a language independent approach in development of our system. Although our system is based on a POS tagger like AMIRA, but the NER subtask in the system is language independent. Also There are many methods to generate POS tags with language independent approaches. Anyway, our method could adopt itself to any other languages which have an accurate POS tagger software.

Next generation of this system can be developed by using more feature functions and predicates which are created specially for Arabic language. Also we can add extracting other types of named entities to this system. For this cases, we need to make special gazetteers for names of locations and organizations.

As we mentioned in section 2., some of other systems for Arabic NER task, use hybrid models like combining multiple CRF models or even multiple methods to improve the results. Using such approaches can improve our system too.

7. References

H. Al-Jumaily, P. Martínez, J.L. Martínez-Fernández, and E. Van der Goot. 2011. A real time Named Entity Recognition system for Arabic text mining. *Language Resources and Evaluation*, pages 1–21.

Y. Benajiba and P. Rosso. 2007. ANERsys 2.0: Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information.

In *Proc. of Workshop on Natural Language-Independent Engineering, IICAI-2007*.

Y. Benajiba and P. Rosso. 2008. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153.

Yassine Benajiba, Mona Diab, and Paolo Rosso. 2004. ARABIC NAMED ENTITY RECOGNITION: AN SVM APPROACH.

Y. Benajiba, P. Rosso, and J. BenedíRuiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. *Computational Linguistics and Intelligent Text Processing*, pages 143–153.

Mona Diab. 2009. Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. In Khalid Choukri and Bente Maegaard, editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April. The MEDAR Consortium.

Charles Elkan. 2008. Log-linear models and conditional random fields.

A. Elsebai and F. Meziane. 2011. Extracting person names from Arabic newspapers. In *Innovations in Information Technology (IIT), 2011 International Conference on*, pages 87–89. IEEE.

H. Fehri, K. Haddar, and A.B. Hamadou. 2011. Recognition and Translation of Arabic Named Entities with NooJ Using a New Representation Model. In *International Workshop Finite State Methods and Natural Language Processing*, page 134.

Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. On Arabic Transliteration. In Abdelhadi Soudi, Antal van den Bosch, Günter Neumann, and Nancy Ide, editors, *Arabic Computational Morphology*, volume 38 of *Text, Speech and Language Technology*, pages 15–22. Springer Netherlands.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Eighteenth International Conference on Machine Learning*.

Charles Sutton and Andrew McCallum. 2007. An Introduction to Conditional Random Fields for Relational Learning.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Advanced Search in Quran: Classification and Proposition of All Possible Features

Assem Chelli,Amar Balla,Taha zerrouki

ESI Ecole nationale Superiure d'Informatique
Algiers
assem_ch@gmail.com, a_balla@esi.dz, taha.zerrouki@gmail.com

Abstract

This paper contains a listing for all search features in Quran that we have collected and a classification depending on the nature of each feature. it's the first step to design an information retrieval system that fits to the specific needs of the Quran

Keywords: Information Retrieval, Quran, Search Features

1. Introduction

Quran,in Arabic, means the read or the recitation. Muslim scholars define it by: the words of Allah revealed to His Prophet Muhammad, written in mushaf and transmitted by successive generations (التواتر) (Mahssin, 1973). The Qur'an is also known by other names such as: Al-Furkān , Al-kitāb , Al-dhikr , Al-wahy and Al-rōuh . It is the sacred book of Muslims and the first reference to Islamic law.

Due to the large amount of information held in the Qur'an, it has become extremely difficult for regular search engines to successfully extract key information. For example, When searching for a book a book related to English grammar, you'll simply Google it, select a PDF and download it .that's all! Search engines (like Google) are utilized generally on Latin letters and for searching general information of document like content, title, author...etc. However, searching through Qu'ranic text is a much more complex procedure requiring a much more in depth solution as there is a lot of information that needs to be extracted to fulfill Quran scholar's needs. Before the creation of computer, Quran scholars were using printed lexicons made manually. The printed lexicons can't help much since many search process waste the time and the force of the searcher. Each lexicon is written to reply to a specific query which is generally simple. Nowadays, there are applications that are specific for search needs; most of applications that were developed for Quran had the search feature but in a simply way: sequential search with regular expressions.

The simple search using exact query doesn't offer better options and still inefficient to move toward Thematic search by example. Full text search is the new approach of search that replaced the sequential search and which is used in search engines. Unfortunately, this approach isn't applied yet on Quran. The question is why we need this approach? Why search engines? Do applications of seach in Quran really need to be implemented as search engines? The features of search that we'll mention them in this paper will answer those questions.

Our proposal is about design a retrieval system that fit the Quran search needs. But to realize this objective, we must first list and classify all the search features that are possible and helpful. We wrote this paper to explain this point. The paper contains a listing for all search features that we have collected and a classification depending on the nature of feature and the way how can be implemented.

We'll go through the problematic for a start. Secondly, we'll set out an initial classification and list all the search features that we think that they are possible.

2. Problematic

To clarify the vision about the problematic of this paper, we are describing the challenges that face the search in Quran:

- First, as a general search need;
- Second, as an Arabic search challenge;
- Third, Quran as a special source of information.

We start explaining the first point, the search in Quran is by theory has the same challenges of search in any other documents. The search in documents has passed by different phases in its evolution. At the beginning, the search was sequential based an exact keyword before the regular expressions were introduced. Full text search was invented to avoid the sequential search limitations on huge documents. The full text search introduces some new mechanisms for text analysis that include tokenization, normalization, and stemming...etc. Gathering Statistics make now a part of search process, it helps to improve the order of results and the suggestions. After the raising of the web semantic, the search is heading to a semantic approach where the to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable dataspace to generate more relevant results. To get more user experience, the search engines try to improve the behavior of showing the results by sorting it based on their relevance in the documents , more sorting criteria, Highlighting the keywords, Pagination, Filtering and Expanding. Moreover, improve the query input by introducing different input methods (ex: Vocal input) and suggesting related keywords. Till now, most of these features are not implemented to use on Quran. And many of them need to be customized to fit the Arabic properties that what explained in the next point.

Secondly, Quran's language is considered as the classical Arabic.Arabic is a specific language because its morphology and orthography, and this must be taken into consideration in text analyzing phases. For instance, letters shaping (specially the Hamza -ء-), the vocalization, the different levels of stemming and types of derivations...etc. That must be taken into consideration in search features by example: the regular expressions are badly misrepresenting the Arabic letters since the vocalization diacritics are not distinct from letters. The absence of vocalization issues some ambiguities (Albawwab, 2009) in understanding the words:

- الملك؟ المَلَك, المَلِك، المُلْك
- وعد؟ وَعَدَ ، وَ+عَدَّ
- وله؟ وَلَه، وَلِّ+ه ، وَ+لَ+هُ

Trying to resolve these problems as a generic Arabic problem is really hard since it hasn't linguistic resources to make strict lexical analyzers. By the contrary, Quran has a limited count of words and that means that it's possible to write manually morphological indexes and use it to replace lexical analyzers. Finally, we explain in this point what the specific challenges faced in search in order of the particular characteristic of Quran. El-Mus-haf , the book of Quran, is written on the Uthmani script . This last is full of recitation marks and spells some words in a different way than the standard way. By example, the word "بسطة" is spelled "بصطة" in Uthmani. The Uthmani script requires considering its specifications in Text analyzing phases: Normalization, Stemming. Quran is structured in many analytic levels (BenJammaa, 2009):

- Main structure: Sura, Aya, word, letter.
- Special emplacements: Sura's start, Sura's end. Aya's end. Sajdah, Waqf , Facilah;
- Mushaf structure: page, thumn, rubu', nisf, hizb, juz';
- Quran writing: sawāmit, harakāt, hamza, diacritics, signes of distinction between similar letters, phonetic signs;
- Incorporeal structure: word, keyword, expression, objectif unit
- Revelation: order, place, calender, cause, context...etc.

The users may need to search, filter results or group them based on one of those structures. There are many sciences related to Quran, named Quranic Sciences: Tafsir, Translation, Recitation, Similitude and Abrogation...etc.

Next, we'll propose an initial classification for the search features that we have inspired from the problematic points.

3. Classification

To make the listing of search features easier, we classified them in many classes based on their objectives.

1. **Advanced Query**: This class contains the modifications on simple Query in order to give the user the ability of formulating his query in a précised way. By example: Phrase search, Logical relations, Jokers.
2. **Output Improvement**: Those are to improve the results before showing it to users. The results must pass by many phases: Scoring, Sorting, Pagination, Highlighting...etc.
3. **Suggestion Systems**: This class contains all options that aims to offer a suggestion that help users to correct, extend the results by improving the queries. By example, suggest correction of misspelled keywords or suggest relative-words.
4. **Linguistic Aspects**: This is about all features that are related to linguistic Aspects like stemming, Selection&filtring stop words, normalization.
5. **Quranic Options**: It's related to the properties of the book and the information included inside. As we mentioned in the problematic, the book of Quran (al-mushaf) is written in uthmani script full of diacritization symbols and structured in many ways.
6. **Semantic Queries**: Semantic approach is about to allow the users to pose their queries in natural language to get more relevant results implicitly.
7. **Statistical System**: This class covers all the statistical needs of users. By example, searching the most frequented word.

This is an initial classification; we have to improve it for a well exploit of all possible search features.

4. Proposals

In this point, we enlist all possible search features based on the classification we mentioned before. These entire features express a search need: general, related to Arabic or related to Quran. We have collected the basic ideas from:

- Classic & Semantic search engines: Google,
- Arabic search engines: Taya it,
- Quranic search tools: Zekr application, almonaqeb alqurany,
- Indexing/Search programming libraries: Whoosh, Lucene
- Quranic Paper lexicons: The indexed mu'jam of words of the Holy Quran(المعجم المفهرس لألفاظ القرآن الكريم) by Mohammed Fouad Abd El-bāki

We have manipulated those ideas to fit the context of Arabic and Quran. There are many features that are totally new , we propose them to fulfill a search need or resolve a specific problem. In addition to simple search, these are our proposals:

1. **Advanced Query**

   (a) **Fielded search**: uses fieldname in the query to search in a specific field. Helpful to search more information like surah names.

   - سورة:الفاتحة

   (b) **Logical relations**: to force the existence or absence of a keyword. The most known relations are: AND for conjunction, OR for disjunction and NOT for exception. The relations can be grouped using parenthesis.

   - (الصلاة - الزكاة) + سورة:البقرة

   (c) **Phrase search**: is a type of search that allows users to search for documents containing an exact sentence or phrase.

   - ”الحمد لله“

   (d) **Interval search:** used to search an interval of values in the numerical field. Helpful in fields like: Ayah ID, Page, Hizb, statistic fields.

   - رقم_الآية : [1 إلى 5]

   (e) **Regular expressions (Jokers):** used to search a set of words that share some letters. This feature can be used to search a part of a word. In latin , there is two Jokers used largely : ? replaces one letter, * replaces an undefined number of letters. These jokers are inefficient in Arabic because the existence of vocalization symbols which they are not letters, and Hamza(ء) letter that has different forms (different Unicode emplacements) .

   - ب؟طة = بسطة، بصطة
   - *نبي* = نبي ، النبيين ، الأنبياء . . .

   (f) **Boosting:** used to boost the relevance factor of any keywords.

   - سميع^2 بصير

   (g) **Combining features:** search using a combination of the previous elementary features.

   - ”*حمد لله“^2

2. **Output Improvments**

   (a) **Pagination:** dividing the results on pages.

   - 10, 20,50... results per page

   (b) **Sorting**: sort the results by various criteria such as:

   - Score
   - Mushaf order
   - Revelation order
   - Numerical order of numeric fields
   - Alphabitical or Abjad order of alphabetic fields
   - A combination of the previous orders

   for the alphabetical order ,we must consider the real order of:

   - Hamza forms : أ ء ئ ؤ
   - Ta' forms: ة ت
   - Alef forms : ا ى

   (c) **Highlight:** used for distinction of searched keywords in Ayah.

   - الحمد <style>لله</style>رب العالمين

   (d) **Real time output**: used to avoid the time of user waiting ,and show the results directly when retrieved.

   (e) **Results grouping:** the criteria can be used for grouping are:

   - by similar Ayahs
   - by Surahs
   - by subjects
   - by taffsir dependency
   - by revelation events
   - by Quranic examples

   (f) **Uthmani script with full diacritical marks:**

   - ۞ لَّقَدْ كَانَ فِى يُوسُفَ وَإِخْوَتِهِۦٓ ءَايَٰتٌ لِّلسَّآئِلِينَ]]

3. **Suggestion Systems**

   (a) **Vocalized spell correction**: offers alternatives for keywords when misspelled or appeared with a different form in Quran.

- أبرَاهَام: إبْرَاهِيم

(b) **Semantically related keywords (Ontology-Based)** : used as hints, generally based on domain ontologies

- يعقوب: يوسف، الأسباط، نبي...

(c) **Different vocalization**: suggest all the possible vocalizations sorted by the frequencies.

- الملك : المَلِك ، المُلْك، المَلَك ...

(d) **Collocated words**: provides a completion based on the collocation statistics.

- سميع : سميع عليم، سميع بصير
- الحمد: الحمد لله

(e) **Different significations:** used to limit a keyword on only one of their meanings.

- رب: معنى 1 (إله)، معنى 2 (سيد)

4. **Linguistic Aspects**

(a) **Partial vocalization search:** gives user the opportunity to specify some diacritics and not all.

- مَلـك_ to find مَلِك, مَلَك ... and ignore مُلْك

(b) **Multi-level derivation**: the Arabic words derivations can be devided to four levels : exact word فاسقيناكموه , Word affixes removed (Lemma) أسقينا , Stem: أسقى , Root : سقي .

- (word:سقي , level: root) to find يَسْقُونَ, نَسْقِي, وَيَسْقِينِ , يُسْقَوْنَ , وَيُسْقَوْنَ, وَسُقْيَاهَا , وَأَسْقَيْنَاكُمْ , وَيُسْقَى , لَأَسْقَيْنَاهُمْ , وَنُسْقِيهُ , فَأَسْقَيْنَاكُمُوهُ , فَسَقَى , السِّقَايَةَ, نُسْقِيكُمْ, فَيَسْقِي, سِقَايَةَ, تَسْقِي, سَقَيْتَ, يُسْقَى, اسْتَسْقَى, اسْتَسْقَاهُ, وَسُقُوا.
- (Word: أسقينا, level: lemma) to find وَأَسْقَيْنَاكُمْ , لَأَسْقَيْنَاهُمْ, فَأَسْقَيْنَاكُمُوهُ .

(c) **Specific-derivations**: this is specification of the pervious feature .Since Arabic is fully flexional , it has lot of derivation operations .

- Conjugation in Past tense of قال to find قالت, قال, قالوا, قلن ...
- Conjugation in Imperative tense de قال to find قل , قولوا ...

(d) **Word properties embedded query:** offers a smart way to handle words families by filtering using a set of properties like: root , lemma, type, part-of-speech, verb mood, verb form, gender, person, number, voice...etc.

- { جذر:ملك، نوع:اسم، عدد: مفرد }

(e) **Numerical values substitution:** this helps to retrieve numbers, even if appeared as words.

- 309 replaces ثلاثمائة وتسعة

(f) **Consideration/Ignoring spell faults:** especially for the letters that usually misspelled like Hamza(ء); The Hamza letter is hard to write its shape since its writing is based on the vocalization of both of it and its antecedent.

- مءصدة replaces مؤصدة
- ضحي replaces ضحى
- جنه replaces جنة

(g) **Uthmani writing way:** offers the user the ability of writing words not as it appeared in Uthmani Mushaf .

- بصطة replaces بسطة
- نعمت replaces نعمة

(h) **Resolving pronoun references:** Pronouns usually refer to other words, called their antecedents because they (should) come before the pronoun. This feature gives the oppurtunity of search using the antecedents.

- لا اله إلا هو, هو ش الله

5. **Quranic Options**

(a) **Recitations marks retrieving:** helpful for Tajweed scholars.

- سجدة : نعم
- نوع _ سكتة:واجبة
- قلقلة:نعم

(b) **Structural option**: since Quran is devided to parts () and the part to Hizbs and Hizb to Halfs ... till we get Ayas.There is also other structures of Quran such as to pages ...etc.

- صفحة: 1
- حزب: 60

(a) **Translation embedded search:** helps users to search using words translations to other langugaes instead of the native arabic text.

- { text: mercy , lang: english , author: shekir }

(b) **Similitudes search (المتشابهات)**

- متشابهة _ {55,13}
[فَبِأَيِّ آلَاءِ رَبِّكُمَا تُكَذِّبَانِ] , 31 exact similitudes

(c) **Examples search (الأمثال)**

- مثل في سورة:البقرة
[مَثَلُهُمْ كَمَثَلِ الَّذِي اسْتَوْقَدَ نَارًا فَلَمَّا أَضَاءَتْ مَا حَوْلَهُ ذَهَبَ اللَّهُ بِنُورِهِمْ وَتَرَكَهُمْ فِي ظُلُمَاتٍ لَا يُبْصِرُونَ]

6. **Semantic Queries**

   (a) **Semantically related words**: there are several different kinds of semantic relations: Synonymy, Antonymy, Hypernym, Hyponymy, Meronymy, Holonymy, Troponymy.

   - Syn(جنة) to search for جنة, نعيم, فردوس ...
   - Ant(جنة) to search for جحيم, سعير , جهنم, سقر ...
   - Hypo(جنة) to search for فردوس ، عدن ...
   - ...

   (b) **Natural questions:** this means to offer the option of forming the search query as a form of an Arabic natural question . the most used Arabic question words are: هل؟ من؟ ما؟ أين؟ متى؟ لم؟ كم؟ أيّ؟ لمن؟

   - من هم الأنبياء؟
     (Who are the prophets?)
     [إِنَّا أَوْحَيْنَا إِلَيْكَ كَمَا أَوْحَيْنَا إِلَى نُوحٍ وَالنَّبِيِّينَ مِنْ بَعْدِهِ وَأَوْحَيْنَا إِلَى إِبْرَاهِيمَ وَإِسْمَاعِيلَ وَإِسْحَاقَ وَيَعْقُوبَ وَالْأَسْبَاطِ وَعِيسَى وَأَيُّوبَ وَيُونُسَ وَهَارُونَ وَسُلَيْمَانَ وَآتَيْنَا دَاوُودَ زَبُورًا] - النساء 163
   - ما هي الحطمة؟
     (What is Al-hottamat?)
     ؟ [نَارُ اللَّهِ الْمُوقَدَةُ] - الهمزة 6
   - أين غلبت/هزمت الروم؟
     (Where was Rome defeated?)
     [فِي أَدْنَى الْأَرْضِ وَهُمْ مِنْ بَعْدِ غَلَبِهِمْ سَيَغْلِبُونَ] - الروم 3
   - كم مكث أصحاب الكهف؟
     (How much of time did People of Cave stay?)
     [وَلَبِثُوا فِي كَهْفِهِمْ ثَلَاثَ مِائَةٍ سِنِينَ وَازْدَادُوا تِسْعًا] - الكهف 25
   - متى يوم القيامة؟
     (When is the Day of Resurrection?)
     [يَسْأَلُكَ النَّاسُ عَنِ السَّاعَةِ قُلْ إِنَّمَا عِلْمُهَا عِنْدَ اللَّهِ وَمَا يُدْرِيكَ لَعَلَّ السَّاعَةَ تَكُونُ قَرِيبًا] - الكهف 25
   - كيف يتشكّل الجنين؟
     (How has the embryo be formed?)
     [ثُمَّ خَلَقْنَا النُّطْفَةَ عَلَقَةً فَخَلَقْنَا الْعَلَقَةَ مُضْغَةً فَخَلَقْنَا الْمُضْغَةَ عِظَامًا فَكَسَوْنَا الْعِظَامَ لَحْمًا ثُمَّ أَنْشَأْنَاهُ خَلْقًا آخَرَ فَتَبَارَكَ اللَّهُ أَحْسَنُ الْخَالِقِينَ] - المؤمنون 14

   (c) **Automatic diacritization :** the absence of diacritics lead to the ambiguities that we've mentioned them in Problematic. This feature helps to pass over these ambiguities and try to resolve them before executing the search process.

   - رسول من الله == رَسُول مِنَ اللهِ

   (d) **Proper nouns search**: lot of proper nouns are mentions clearly in Quran but some of them are just referred to implecitly . This feature is to search for proper nouns however mentioned : explicitly or implicitly.

   - بنيامين؟
     [إِذْ قَالُوا لَيُوسُفُ وَأَخُوهُ أَحَبُّ إِلَى أَبِينَا مِنَّا وَنَحْنُ عُصْبَةٌ إِنَّ أَبَانَا لَفِي ضَلَالٍ مُبِينٍ] - المؤمنون 14
   - أبو بكر/ الصديق؟
     [ثَانِيَ اثْنَيْنِ إِذْ هُمَا فِي الْغَارِ إِذْ يَقُولُ لِصَاحِبِهِ لَا تَحْزَنْ إِنَّ اللَّهَ مَعَنَا] - التوبة 40

7. **Statistical System**

   (a) **Unvocalized word frequency:** this feature is about gathering the frequency of each word per ayahs ,per surahs , per hizbs.

   - How many words of “الله” in Sura المجادلة?
   - What the first ten words which are the most frequently cited words in the whole Quran?

   (b) **Vocalized word frequency:** the same as the previos feature but with consideration of diacritics.that will make the difference for instance between مِنْ, مَنْ and مَنًّ .

   (c) **Root/Stem/Lemma frequency:** the same as the previous feature but use Root, Stem or Lemma as unit of statistics instead of the vocalized word.

   - How many the word of “بحر” and its derivations are mentioned in the whole Quran? the word “بحار” will be considered also.

   (d) **Another Quranic units frequency:** the statistics could be gathered based on many units otherwise the words like letters,ayas, Recitation marks...etc.

   - How many letters in the Sura “طه”?
   - What's the longest Ayah?
   - How many Marks of Sajda in the whole Quran?

5. Related Works

There is lot of web applications and softwares that offer the service of search in Quran but most of them are simple search using exact words. We'll mention some applications and websites that implemented some special search features.

Alawfa is a website that offers search in Quran and Hadith. He has a fast retrieving method. It offers the highlight, pagination. The main disadvantage of Alawfa is that it doesn't respect the specifics of Arabic and Quran. It searches only as exact word or as a part of word.(Ala, 2011)

Al-Monaqeb-Alqurany is a search service of Quran included in the website www.holyquran.net. It offers a board for advanced search options that contains many useful options(Alm, 2011) :

- Transliteration
- Ignore/Consider Hamza spelling errors
- Limit the search in an interval of ayahs and surahs
- Search using Arabic words root
- Highlight & Pagination

Zekr is a desktop application aimed to be an open platform Quran study tool for simply browsing and researching on Quran and its translations. It offers many sorting criteria like Relevance, Natural order, Revelation order, Aya length. Also, it offers browsing with Mushaf Structures: Hizb number, Surah name,. It search using the exact word, the part of word, the arabic root (Zek, 2012).

6. Conclusion

In this paper, we have enlisted the search features in Quran that it's helpful. To facilitate the enlisting, we have classified those features depending on the nature of the problem. That list will help us to make a detailed retrieval system that fit perfectly the needs of the Quran. Each feature has a different level of complexity: could be implemented easily or may lead to vast problem that need a deeper study.

References

2011. Alawfa website.

Merouane Albawwab. 2009. محركات_البحث_في_النصوص_العربية_وصفحات_الأنترنت . مجمع_اللغة_العربية_دمشق .

2011. Al-monakkeb al-corani website.

Muhammed BenJammaa. 2009. منهجية_تعاونية_لإنجاز_موسوعة_إلكترونية_شاملة_للقرآن الكريم_وعلومه.

Muhammed Salem Mahssin. 1973. .دار_الأصفهاني_للطباعة_بجدة .تاريخ_القرآن_الكريم

2012. Zekr - quran study software website.

An Introduction to Noor Diacritized Corpus

[1]Akbar Dastani, [2]Behrouz Minaei-Bidgoli, [3]Mohammad Reza Vafaei, [4]Hossein Juzi
[1234]Computer Research Center of Islamic Sciences, Qom, Iran
[2]Iran University of Science and Technology, Tehran, Iran

E-mail: a-dastani@noornet.net, b_minaei@iust.ac.ir, vafaie@noornet.net, hjuzi@noornet.net

Abstract

This article is aimed to introduce Noor Diacritized Corpus which includes 28 million words extracted from about 360 hadith books. Despite lots of attempts to diacritize the holy Quran, little diacritizing efforts have been done about hadith texts. This corpus is therefore from a great significance. Different statistical aspects of the corpus are explained in this article. This paper states challenges of diacritizing activities in Arabic language in addition to general specifications of the corpus.

Keywords: Noor Diacritized Corpus, diacritization, Arabic corpora

1. Preface

Building diacritized corpora for developing language-oriented investigations is of great necessity in Arabic language since its broad vocabulary and word forms and meanings. It can be used to a large extent in linguistic studies, grammar, semantics, natural language processing (NLP), information Retrieval (IR) and many other fields. Many centers working on NLP and IR are faced with a restriction on accessibility to these corpora.

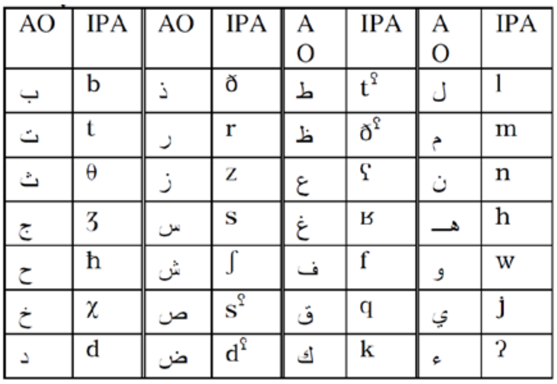
Computer Research Center of Islamic Sciences (CRCIS) has undertaken diacritization of hadith texts with the intention of facilitating the public's use of hadiths, clarifying the concepts of texts and protecting the rich content of Islamic scriptures.

2. Diacritization

The word correctness does not mean only spelling correctness because there is a more extensive definition for the correctness; that is, presenting the word in a way that represents the required concept. Therefore, this requires the employment of special signs so-called "diacritic".

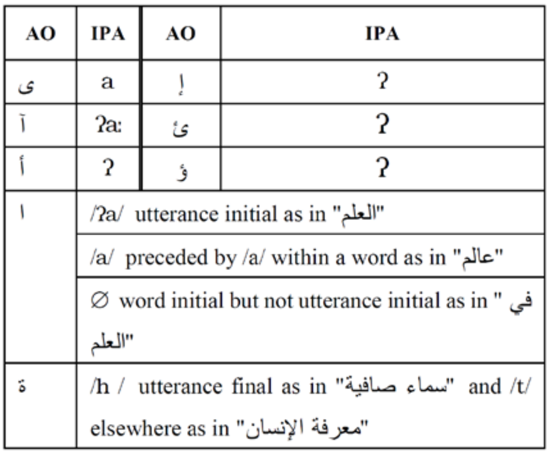
Diacritization is a science that studies the letters' diacritical marks, including short diacritics and long diacritics as well as Sukoun (absence of vowel), Tashdid (germination of consonants) and Tanwin (adding final n sound).

Similar to any other languages, phonetic system of Arabic consists of consonants and vowels. It has three short vowels and three long ones. There are 28 letters (consonants) in Arabic (table 1) (Alghamdi et al, 2003). The writing system of Arabic consists of 36 consonants (table 1, 2) and 8 vowels (table 3). When fully diacritizing an Arabic text, every letter should have a mark, except for some letters in special combinations (table 4) (Alghamdi et al, 2007).



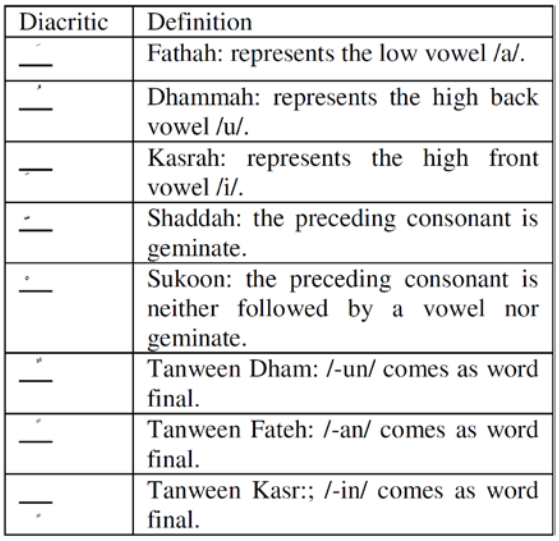
AO	IPA	AO	IPA	A O	IPA	A O	IPA
ب	b	ذ	ð	ط	$t^{ʕ}$	ل	l
ت	t	ر	r	ظ	$ð^{ʕ}$	م	m
ث	θ	ز	z	ع	ʕ	ن	n
ج	ʒ	س	s	غ	ʁ	هـ	h
ح	ħ	ش	ʃ	ف	f	و	w
خ	χ	ص	$s^{ʕ}$	ق	q	ي	j
د	d	ض	$d^{ʕ}$	ك	k	ء	ʔ

Table 1: Arabic Orthography (AO) and their representations in International Phonetic Alphabet (IPA) as the consonant inventory of Modern Standard Arabic (Alghamdi et al, 2003).



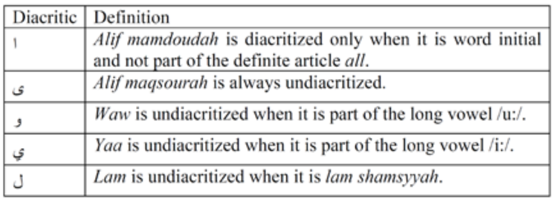
<table>
<tr><th>AO</th><th>IPA</th><th>AO</th><th>IPA</th></tr>
<tr><td>ى</td><td>a</td><td>إ</td><td>ʔ</td></tr>
<tr><td>آ</td><td>ʔa:</td><td>ئ</td><td>ʔ</td></tr>
<tr><td>أ</td><td>ʔ</td><td>ؤ</td><td>ʔ</td></tr>
<tr><td rowspan="3">ا</td><td colspan="3">/ʔa/ utterance initial as in "العلم"</td></tr>
<tr><td colspan="3">/a/ preceded by /a/ within a word as in "عالم"</td></tr>
<tr><td colspan="3">∅ word initial but not utterance initial as in " في العلم"</td></tr>
<tr><td>ة</td><td colspan="3">/h / utterance final as in "سماء صافية" and /t/ elsewhere as in "معرفة الإنسان"</td></tr>
</table>

Table 2: Additional Arabic orthographic symbols (AO) and their IPA representations (Alghamdi et al, 2003).



Diacritic	Definition
ـَ	Fathah: represents the low vowel /a/.
ـُ	Dhammah: represents the high back vowel /u/.
ـِ	Kasrah: represents the high front vowel /i/.
ـّ	Shaddah: the preceding consonant is geminate.
ـْ	Sukoon: the preceding consonant is neither followed by a vowel nor geminate.
ـٌ	Tanween Dham: /-un/ comes as word final.
ـً	Tanween Fateh: /-an/ comes as word final.
ـٍ	Tanween Kasr:; /-in/ comes as word final.

Table 3: Arabic diacritics. The horizontal line represents an Arabic letter (Alghamdi et al, 2003).



Diacritic	Definition
ا	*Alif mamdoudah* is diacritized only when it is word initial and not part of the definite article *all*.
ى	*Alif maqsourah* is always undiacritized.
و	*Waw* is undiacritized when it is part of the long vowel /u:/.
ي	*Yaa* is undiacritized when it is part of the long vowel /i:/.
ل	*Lam* is undiacritized when it is *lam shamsyyah*.

Table 4: Arabic letters that are not diacritized (Alghamdi et al, 2007).

2.1. Challenges in Diacritizing

There are various challenges ahead of building diacritized corpora of Arabic texts, some of which are mentioned hereby:

- **Interactivity of adjacent letters' diacritics:** Normally, diacritic marks of the non-ending letters of a word are not changed, but the ending letters of mu`rab nouns and verbs change in accordance with the position they take in the sentence; this is the subject of syntax (Naḥw). In such cases, the researcher should take into account the surrounding words. Let's take the word "عبدالله" as an example. If the first word is pronounced "`Abada" (عَبَدَ) then the latter will be pronounced "Allâha" (اللهَ). If it is "`Ubida" (عُبِدَ) , the latter is "Allâhu" (اللهُ), etc.
- **Diacritic of the words sanctioned by audition:** Most of Arabic words are regular; therefore, one who knows the grammatical rules of Arabic can recognize the diacritic marks of words. There are, however, many words in Arabic, like other languages, that do not follow the rules. Such words are so-called "Simâ`î" (auditive) in Arabic. Although the diacritic mark of auditive words can be determined by looking them up in a dictionary, there are some words remained whose diacritics cannot be easily recognized, e.g. "السقمونيا".
- **Basic Disagreements:** There are basic disagreements in Arabic diacritization, some examples of which are the following:
  - For conjunctive hamza, some editors prefer not to put a mark on it, while some others keep it intact although they know it has no role in pronunciation. Others, like Uthman Taha, put a special mark (similar to ص) on it.
  - Due to difficulty of applying all the rules about hamza, there are differences witnessed in recording the words containing hamza. For example the three forms of "تشاءون", "تشاؤن" and "تشاؤون" represent a single word.
- **Neighboring Consonants**: No one (in Arabic) can pronounce two neighboring consonants if none is followed by a vowel.
- **Difficult Words:** Rarely used words whose diacritics seem strange.
- **Multi-Dimensional Words:** Some words have acceptable meanings with different pronunciations. This is more controversial when it comes to recitation of the Quran. For example, the phrase "عبد الطاغوت" can be pronounced either as "`Abadat-Tâğût" or "`Ubidat- Tâğût". About the disagreement in pronouncing hadith words, we do not have sufficient sources in hand. In other words, this matter has been subjected to study in recent years.

Having comprehended necessity of diacritizing Islamic hadith texts, Computer Research Center of Islamic Sciences (CRCIS) has intended to undertake this task.

3. Corpus Introduction

The CRCIS commenced diacritizing Islamic hadith texts with the book "Wasa'il al-Shi`ah" in the beginning of the recent decade with the intention of safeguarding Islamic rich texts and using them in its computer dictionaries and software products. Despite lots of attempts to diacritize the holy Quran, little diacritizing efforts have been done about hadith texts.

In the first phase, diacritizing was done by the experts in this field on printed books. The output of the first phase was revised by a second group. Then, the diacritized texts were entered into the computer. Regarding the problems of physical diacritization, the task was thereafter done directly in the computer by the experts.

When the corpus had grown to a considerable extent, it was the time to design a program to diacritize more quickly, more exactly and automatically based on the similar works done manually. Then the results of machine diacritization, have verified by experts (Verification time for 500 pages was about 200 person-hour, in average). As a result of this continual endeavor, 135 books in 358 volumes including Kutub al-Arba`a, Wasa'il al-Shi`ah, Bihar al-Anwar and Mustadrak al-Wasa'il were diacritized.

4. Statistical Properties of Corpus

There are 28,413,788 diacritized words in the corpus, 639,243 unique words. Furthermore, there are 339,684 unique non-diacritized words in the corpus.

Table 5 shows the frequency of words with specific number of diacritization forms (up to 10 forms). For example, 212,938 words in the corpus have only one diacritization form.

Each word has different diacritization forms according to linguistic rules. The words shown in table 6 have the largest number of diacritization forms.

number of diacritization forms	frequency of words
1	212,938
2	59,978
3	29,630
4	14,892
5	8,057
6	4,836
7	3,140
8	1,927
9	1,190
10	861

Table 5: the number of diacritization forms of words.

حَجَرٌ	فَرَقٌ	إِذَنِ	تَعَلَّمَ	مَحَلَ	كَذَبَةٌ
حَجَرٍ	فَرَقٍ	إِذَنْ	تَعَلَّمُ	مَحَلٌّ	كَذَبَتِ
حَجَرَ	فَرَقَ	إِذْنٌ	تَعَلَّمِ	مَحَلٍّ	كَذَبَتْ
حَجَرُ	فَرَقُ	إِذْنٍ	تَعَلَّمْ	مَحَلَّ	كَذَبْتَ
حَجَرِ	فَرَقِ	إِذْنَ	تَعَلُّمٌ	مَحَلُّ	كَذَبْتُ
حَجِرٍ	فَرَقَّ	إِذْنُ	تَعَلُّمٍ	مَحَلِّ	كَذَبْتِ
حَجْرٌ	فَرُقَ	إِذْنِ	تَعَلُّمَ	مَحَلّ	كَذِبَةً
حَجْرٍ	فَرُقَّ	إِذَنْ	تَعَلُّمُ	مَحِلَ	كَذِبَةٌ
حَجْرَ	فَرِقَ	إِذْنُ	تَعَلُّمِ	مَحِلُ	كَذِبَةٍ
حَجْرُ	فَرْقٌ	إِذْنٍ	تَعْلَمَ	مَحِلٌّ	كَذِبْتَ
حَجْرِ	فَرْقٍ	إِذْنَ	تَعْلَمُ	مَحِلٍّ	كَذِبْتُ
حَجَّرَ	فَرْقَ	إِذْنُ	تَعْلَمِ	مَحِلَّ	كَذْبَةً
حُجَرٌ	فَرْقُ	إِذْنِ	تَعْلَمْ	مَحِلُّ	كَذْبَةٌ
حُجَرٍ	فَرْقِ	أَذَنِ	تُعَلَّمَ	مَحِلِّ	كَذْبَةٍ
حُجَرَ	فَرَّقَ	أَذِنَ	تُعَلَّمُ	مَحِلّ	كَذَّبَتِ
حُجَرُ	فَرِّقِ	أَذَّنَ	تُعَلَّمْ	مَحْلُ	كَذَّبَتْ
حُجَرِ	فَرِّقْ	أَذِّنْ	تُعَلِّمَ	مَحْلِ	كَذَّبْتَ
حُجُرٌ	فُرِقَ	أُذُنٌ	تُعَلِّمُ	مَحَّلَ	كَذَّبْتُ
حُجَرِ	فُرْقٍ	أُذُنٍ	تُعَلِّمْ	مُحَلٌّ	كُذِبَتْ
حُجْرٌ	فُرِّقَ	أُذُنَ	تُعُلِّمَ	مُحَلٍّ	كُذِبْتَ
حُجْرٍ	فِرَقٌ	أُذُنُ	تُعْلَمَ	مُحَلَّ	كُذِبْتُ
حُجْرَ	فِرَقٍ	أُذُنِ	تُعْلَمُ	مُحِلٌّ	كُذِّبَتْ
حُجْرُ	فِرَقَ	أُذِنَ	تُعْلَمْ	مُحِلٍّ	كُذِّبْتَ
حُجْرِ	فِرَقُ	أُذْنٍ	تُعْلِمَ	مُحِلَّ	كُذِّبْتُ
حِجْرٌ	فِرَقِ	أُذْنَ	تُعْلِمُ	مُحِلٌّ	كِذْبَةً
حِجْرٍ	فِرْقٌ	أُذِّنَ	تُعْلِمِ	مُحِلٍّ	كِذْبَةٌ
حِجْرَ	فِرْقٍ	أَذَنْ	تُعْلِمْ	مُحِلّ	كِذْبَةٍ
حِجْرُ	فِرْقِ				
حِجْرِ					

Table 6: the largest number of diacritization forms.

In table 7, the number of diacritization forms of words (up to 5 letters) is listed based on their length.

2-letters	Freq.	3-letters	Freq.	4-letters	Freq.	5-letters	Freq.
فِی	703,623	قَالَ	457,094	اللَّهِ	242,362	تَعَالَی	56,728
عَنْ	517,041	عَلَی	327,023	عَلَیْهِ	152,910	الْحَسَنِ	37,839
لَا	410,289	کَانَ	162,118	اللَّهُ	131,489	عَلَیْهَا	19,706
مِنْ	410,222	أَبِی	158,504	فَقَالَ	101,963	اللَّهُمَّ	16,411
بْنِ	309,182	ذَلِکَ	124,824	مُحَمَّدِ	66,209	انْتَهَی	15,479
مَا	229,767	عَبْدِ	114,048	مُحَمَّدٍ	57,508	النَّاسِ	15,225
أَنْ	177,739	هَذَا	102,765	الَّذِی	54,589	الْأَرْضِ	14,786
لَمْ	163,187	إِذَا	99,376	لِأَنَّهُ	47,682	النَّبِیِّ	14,282
أَوْ	162,914	أَنَّهُ	90,313	أَحْمَدَ	39,306	عَلَیْهِمْ	12,762

لَهُ	145,890	فِيهِ	85,320	أَبِيهِ	37,238	الَّذِينَ	12,500

Table 7: shows the frequency of top words (sequences of letters) in the CRCIS corpus.

Figure 1 shows the frequency of vowels in the corpus. The vowel “َ” has the highest frequency as it is seen.

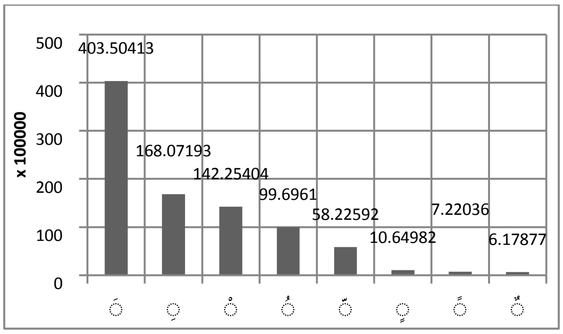


Figure 1: the frequency of vowels in the corpus.

Figure 2 indigates the frequency of letters (consonants) in the corpus. The letter “ل” has the highest frequency holding the number 13,171,176 and the lowest frequency is for the letter “ظ” with the number 198,758.

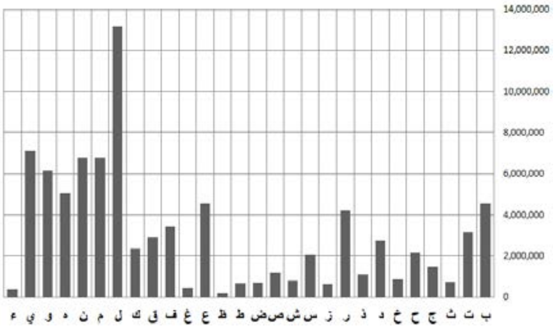


Figure 2: the frequency of Arabic orthography.

The frequency of additional orthographic symbols is shown in figure 3. The letter “ا” (alif) hold the highest frequency number 14,245,061.

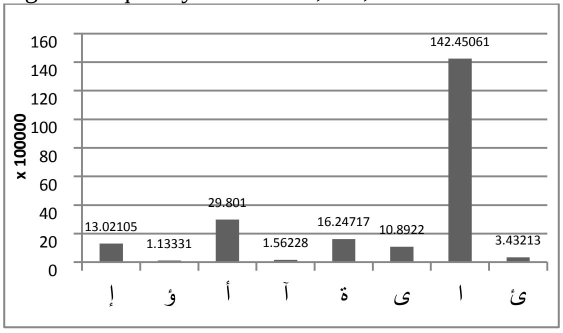


Figure 3: The frequency of additional orthographic symbols.

Figure 4 shows the frequency of distinct words based on their length. 5-letter and 6-letther words are most frequent.

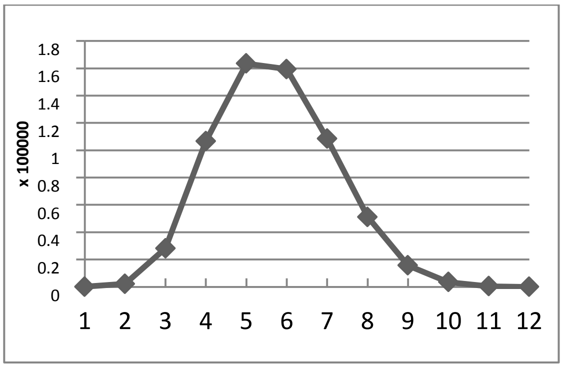


Figure 4: the frequency of distinct n-letter words.

The total of diacritized words is shown in figure 5 based on the length of words. This diagram indicates that 3-letter and 4-letter words are used more frequently despite the fact that there are more 6-letter and 7-letter words.

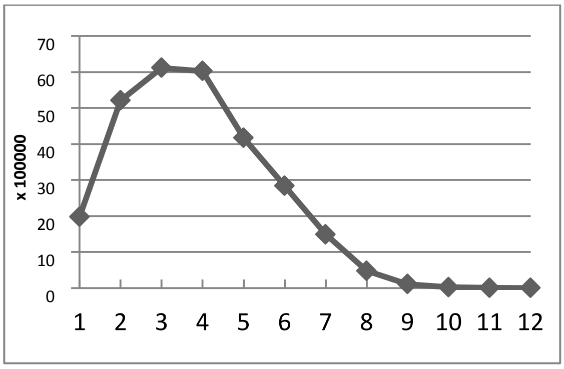


Figure 5: the number of total n-letter words.

Distinct trigrams occurred in the texts amount to 13,998,042 records in the corpus taking into account the diacritics; otherwise, they amount to 13,092,901 records. A sample of trigrams with diacritics is shown in table 8 based on their frequency.

Prev	Current	Next	Freq.
عَبْدِ	اللَّهِ	ع	**63,641**
عَنْ	مُحَمَّدِ	بْنِ	**47,132**
أَبِي	عَبْدِ	اللَّهِ	**43,836**
عَنْ	أَبِي	عَبْدِ	**36,842**
صَلَّى	اللَّهُ	عَلَيْهِ	**30,486**
اللَّهُ	عَلَيْهِ	وَسَلَّمَ	**28,929**
اللَّهِ	ع	قَالَ	**25,979**
عَنْ	أَحْمَدَ	بْنِ	**24,674**
عَنْ	عَلِيٍّ	بْنِ	**23,907**
رَسُولُ	اللَّهِ	ص	**23,628**

Table 8: shows the frequency of top trigram with diacritic.

Some interesting facts and statistics of this corpus that can be used in such procedures as optimizing intelligent grammatical and syntactic results or intelligent diacritizing engine results include predicting the diacritics of two adjacent words using the final diacritics. There are some examples of final diacritic of two adjunct words in table 9.

Freq.	Prev.	Definition	next	Definition
1,918,785	◌َ	**Fathah**	◌َ	**Fathah**
1,771,942	◌ِ	**Kasrah**	◌َ	**Kasrah**
1,344,053	◌َ	**Fathah**	◌ُ	**Dhammah**
1,283,670	◌ِ	**Fathah**	◌َ	**Fathah**
1,000,944	◌َ	**Fathah**	◌ِ	**Kasrah**

Table 9: the final diacritic of two adjunct words

5. Future Tasks

Following the activities in intelligent processing of texts using this corpus, the CRCIS is determined to: improve the content of the corpus with diacritizing other texts, use the character and word linguistic model of the corpus in intelligent systems like error checker, using the corpus in automatic production of stems, decrease ambiguity of the words in syntactic labeling engine output and optimizing the accidence engine results, using the corpus in producing Noor intelligent diacritization engine using machine learning techniques and artificial intelligence.

6. Where the Corpus Has Been Used?

In its development stage, Noor Diacritized Corpus has been used in such programs as Noor 2.5 and Jami al-Ahadith 3 and 3.5.

7. Acknowledgement

This article has been compiled in Noor Text-Mining Research Center affiliated to Computer Research Center of Islamic Sciences. We extend our grateful thanks to all those who played a role in producing this corpus.

8. References

Alghamdi, M., Alhargan, F., Alkanhal, M., Alkhairi , A., Aldusuqi, M. (2003). Saudi Accented Arabic Voice Bank (SAAVB), Final report, Computer and Electronics Research Institute, King Abdulaziz City for Science and technology, Riyadh, Saudi Arabia.

Alghamdi, M., Muzaffar, Z. (2007). KACST Arabic Diacritizer. The First International Symposium on Computers and Arabic Language.

Some issues on the authorship identification in the Apostles' Epistles

Anca Dinu[1], **Liviu P. Dinu**[2], **Alina Resceanu**[3], **Ion Resceanu**[4]

[1]University of Bucharest, Faculty of Foreign Languages and Literatures, Centre for Computational Linguistics
[2]University of Bucharest, Faculty of Mathematics and Computer Science, Centre for Computational Linguistics
[3]University of Craiova, Faculty of Foreign Languages and Literatures
[4]Archbishopric of Craiova
anca_d_dinu@yahoo.com, ldinu@fmi.unibuc.ro, aresceanu@gmail.com, ionresceanu@yahoo.com

Abstract

The New Testament is a collection of writings having multiple authors. The traditional view, that of the Christian Church, is that all the books were written by apostles (Mathew, Paul, Peter, John) or by the apostles' disciples (Mark and Luke). However, with the development of literary theory and historical research, this point of view is disputed. For example, the authorship of seven out of the fourteen epistles canonically attributed to Paul is questioned by modern scholars: the Pastoral epistles (First Timothy, Second Timothy, and Titus), which are thought to be pseudoepigraphic, another three about which modern scholars are evenly divided (Ephesians, Colossians, Second Thessalonians), and the anonymous Hebrews, which, most scholars agree, wasnt written by Paul, but may have been written by a disciple of Pauls. In this paper we applied two different techniques (PCA and clustering based on rank distance) to investigate the authorship identification of Apostles' Epistles.

Keywords: authorship attribution for religious texts, New Testament, Pauline epistles, rank distance

1. Introduction

1.1. Motivation

The authorship identification problem (Mosteller and Wallace, 1964) is a ancient challenge, and almost in every culture there are a lot of disputed works. The problem of authorship identification is based on the assumption that there are stylistic features that help distinguish the real author from any other possibility (van Halteren et al., 2005).

The text characteristics and parameters used to determine text paternity need not have aesthetic relevance. They must be objective, un-ambiguously identifiable, and quantifiable, such that they can be easily differentiated for different authors.

Literary-linguistic research is limited by the human capacity to analyze and combine a small number of text parameters, to help solve the authorship problem. We can surpass limitation problems using computational and discrete methods, which allow us to explore various text parameters and characteristics and their combinations.

The authorship attribution for religious texts is a very interesting and important problem for every religion, and there are many efforts to investigate them (Jockers et al., 2008), (Sadeghi, 2011), (Koppel et al., 2007).

From the types of techniques used in the study of authorship attribution we will be using stylometry to investigate the epistles in the New Testament whose authorship is disputed among modern scholars, focusing on the Pauline and Petrine epistles.

The New Testament is a collection of writings having multiple authors (Moule, 1971), (von Harnack, 2007), (Milton, 1955). The traditional view, that of the Christian Church, is that all the books were written by apostles (Mathew, Paul, Peter, John) or by the apostles' disciples (Mark and Luke). However, with the development of literary theory and historical research, this point of view is disputed. For example, the authorship of an important number of epistles is canonically attributed to Paul is questioned by modern scholars (Guthrie, 1961), (Milton, 1955): the Pastoral epistles (First Timothy, Second Timothy, and Titus), which are thought to be pseudoepigraphic, another three about which modern scholars are evenly divided (Ephesians, Colossians, Second Thessalonians), and the anonymous Hebrews, which, most scholars agree, wasnt written by Paul, but may have been written by a disciple of Pauls. Another instance where the canonical view is questioned by lay researchers is the Synoptic Gospels (traditionally attributed to Mathew, Mark, and Luke), which have a unique internal relationship, and about which scholars predominantly adopt the Two Source hypothesis, claiming that the gospels by Mathew and Luke followed that of Marks and are inspired by it and a so called Q-source document. The traditional view, however, sees the gospel according to Mathew as the first one to be written and as a possible inspiration for Mark and Luke. The Nag Hammadi apocryphal texts discovered in Egypt in 1945 are also a matter of debate. The only books from the New Testament with undisputed authorship are the seven Pauline epistles that were not mentioned above, while the one epistle whose authorship is generally questioned is Hebrews, these two facts motivating our choice of texts for the experiments. For further details on the religious texts authorship identification we refer to (Koppel et al., 2007).

1.2. On the Pauline epistles

Stylometry was already used in order to analyzed the Pauline epistles and various results has been reported in literature. On the other hand, the Pauline epistles are some of the most investigated works related their paternity(J. Muddiman, 2010), (Barr, 2003).

It is well known that St. Paul's vocabulary is varied. His style is calm and not linear like that of Evangelists, but a living expression of temperament, is dynamic, easily passing from one emotion to another, from one mood to another. Because of this, his style does not seem to take account of the sentence, of its normal and natural conducting flow, of the words that seem not expressive enough, of the unnecessary repetitions, of its brackets that often make us lose the thread of the sentence. However, St. Paul's style is distinguished by the depth and subtlety of the ideas he expressed, being influenced by his extensive, somewhat brilliant thinking. That is why St. Paul can not be considered a storyteller, but an apologist, a polemicist, a true preacher, that easily adapts the overflowing of his words to the subject, to the circumstances from which he took his inspiration and to the readers to whom he addressed. It is the style of a missionary, of a Christian, whose main concern is preaching, not preserving in writing the teaching he embraced. So we come to think that St. Paul is more the one that speaks to us, that stands in front of us. This impression is strengthened by the fact that St. Paul seems to dictate his letters. His disciples Timothy, Silvanus (= Silas) a.o. seem to be his secretaries, sometimes his couriers when needed. The Epistle to the Romans indicates precisely that Tertius was the one who wrote/transcribed the letter (Rom 16.22). Some of the pastoral epistles (I and II Timothy and Titus) seem to have been written by his disciples, who apparently gained a greater freedom of writing. The Epistle to Philemon (Knox, 1959) is among the only ones considered to have been written by Paul himself. All these considerations strengthen the belief that stylometry is yet again a useful research method in this endless thickened dilemma of the style and authorship of the Pauline epistles. Based on these methods, further refined explanations can be brought forward to support older theories and opinions, or simply to open new leads in the research aspect.

Some of the established themes on the style and fatherhood of the Pauline epistles can be found in the concern of those who believed that by using this method they would bring a more suitable contribution to the field. Thus, from this perspective, we can approach the subject of the paternity of the Pastoral Epistles (I and II Timothy and Titus), which are under debate due to St. Paul's ideas against the Gnostic (Wilson, 1958) starting with the School of Tübingen.

On the other hand, the Epistle to the Ephesians is also challenged for its style of circular letter not found in other Pauline epistles (Bruce, 1964). It was constantly compared to the Epistle to the Colossians, considered the matrix upon which the Epistle to the Ephesians was written. Another issue may be represented by the ownership of the Epistle to the Thessalonians II from the perspective of its comparison to the Epistle to the Thessalonians I (Manson, 1962), (Gregson, 1966). Nonetheless, the problem regarding the authorship of the Epistle to the Hebrew (Bruce, 1964), (Hewitt, 1960), (Herring, 1970) should be questioned in direct ratio to the Epistle to the Romans and the analysis of the reporting style of the Pauline Epistles to the Romans I and II, Corinthians, Galatians and other letters, especially the pastoral ones, should be done only in relation to one another. Another aspect that could be approached from this perspective is that of establishing the paternity of the General Epistles (the Catholic Epistles I, II Peter, I, II, III John, James, Jude), their interdependency or the relationship between the Epistles of John and the Revelation and so on.

si n sa se cu o la nu a ce mai din pe un ca ca ma fi care era lui fara ne pentru el ar dar l tot am mi nsa ntr cum cnd
toate al aa dupa pna dect ei nici numai daca eu avea fost le sau spre unde unei atunci mea prin ai att au chiar
cine iar noi sunt acum ale are asta cel fie fiind peste aceasta a cele face fiecare nimeni nca ntre aceasta aceea
acest acesta acestei avut ceea ct da facut noastra poate acestui alte celor cineva catre lor unui alta ati dintre doar
foarte unor va aceste astfel avem aveti cei ci deci este suntem va vom vor de

Table 1: The 120 stopwords extracted as the most frequent words in Romanian language.

1.3. Our approach

The distance/similarity between two texts was measured as the distance between the frequencies with which carefully selected functional words are used in the texts. We have chosen to look at Cornilescu's translations of the New Testament in Romanian, a language for which a list of functional words used in stylometry is far from being settled.
Because in all computational stylistic studies/approaches, a process of comparison of two or more texts is involved, in a way or another, there was always a need for a distance/similarity function to measure similarity (or dissimilarity) of texts from the stylistic point of view. Since our framework used functional words as ordered variables, we used Rank Distance, a metric recent introduced in computational linguistics, with good results in authorship identification(Dinu et al., 2008), (Popescu and Dinu, 2008) and in languages similarity analysis (Dinu and Dinu, 2005). For evaluation, weve tested the distance matrices obtained against the Complete Linkage clustering method.
Our results show that, similarity wise, First Thessalonians pairs with Second Thessalonians, while First Timothy pairs with Titus. Hebrews is classified together with Second Peter and the first Petrine epistle pair with Colossians and Ephesians.

2. Classification experiments

2.1. Principal component analysis

In order to speak of distances, we need to represent the samples (the epistles) as points in a metric space. Using the idea that stopwords frequencies are a significant component of the stylome (Chung and Pennebaker, 2007), we first represented each work as a vector of stopwords frequencies. The stopwords can be seen in table 2.1..
A useful visualisation method is the Principal Components Analysis, which gives us a projection from a high-dimensional space into a low-dimensional one, in this case in 2D. Principal components analysis (PCA) (Duda et al., 2001) is a method of dimensionality reduction. The motivation for performing PCA is often the assumption that directions of high variance will contain more information than directions of low variance. The PCA aims to transform the observed variables to a new set of variables which are uncorrelated and arranged in decreasing order of importance. These new variables, or components, are linear combinations of the original variables, the first few components accounting for most of the variation in the original data. Typically the data are plotted in the space of the first two components.
PCA works in the euclidean space and so implicitly use euclidean distance and standard inner product.
Using this stopword frequency representation, the first principal components plane looks like figure 2.1..

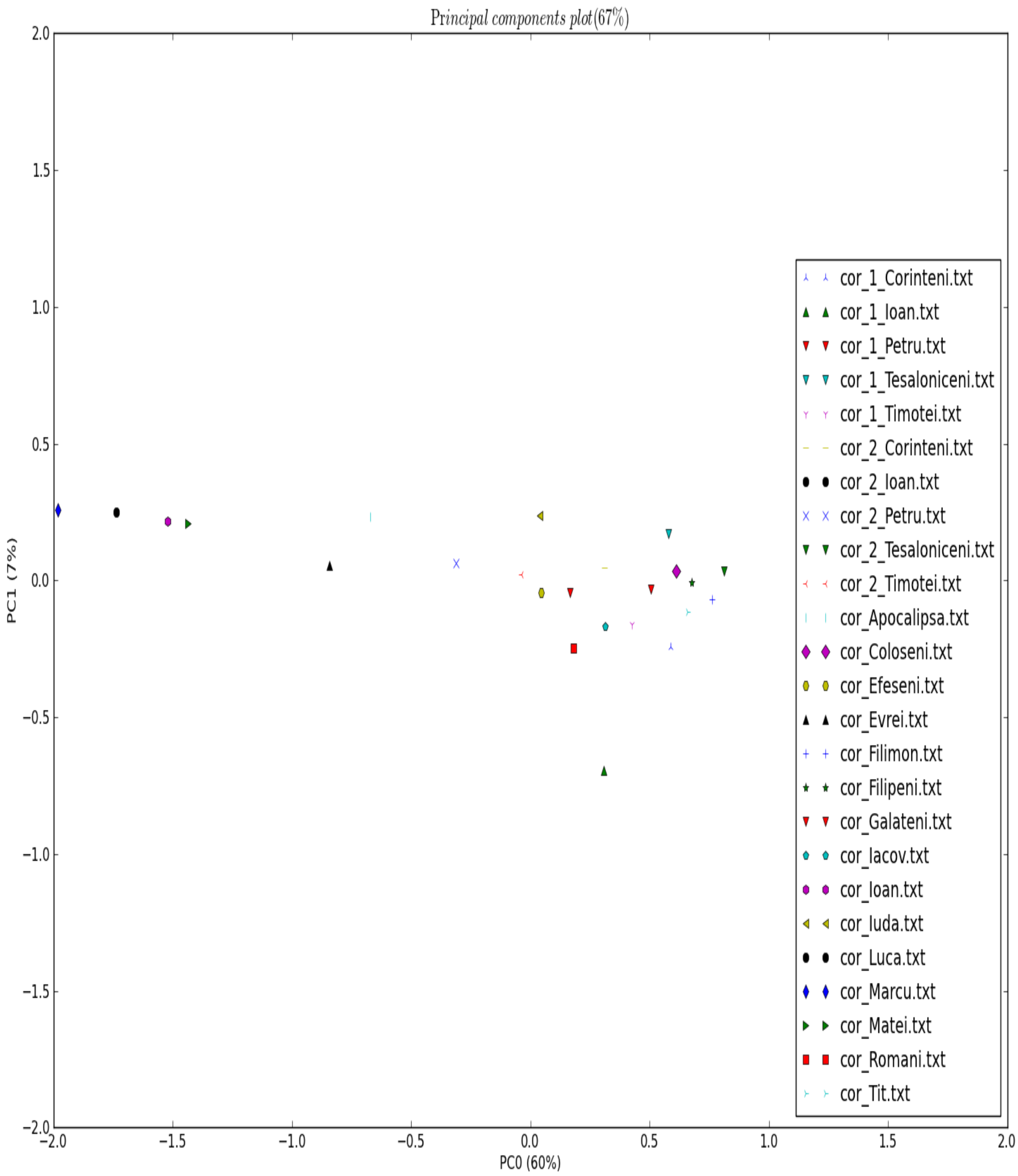


Figure 1: Principal components plot. The cluster on the left consists only of four evangelists. The central cluster is Hebrew. The right cluster is Paul's epistles

We can see in figure 2.1. that some points are well defined. The left cluster consists of the four gospels (Mark, Luke, John and Mathew).

A central point is corresponding to Hebrew; to the right is John (Revelation) and II Peter (which are well delimited by the rest of points).

The right cluster consists of the Pauline epistles and I Peter. The figure 2.1. show a different behavior of Hebrew related to the Pauline corpus; we see also that the Peter II is far from the Peter I, which suggests that the Petrine epistles were written by different authors.

2.2. Clustering experiments

Compared with other machine learning and statistical approaches, clustering was relatively rarely used in stylistic investigations. However, few researchers (Labbe and

Labbe, 2006), (Popescu and Dinu, 2008), (Duda et al., 2001) have recently proved that clustering can be a useful tool in computational stylistic studies.

An agglomerative hierarchical clustering algorithm (Duda et al., 2001) arranges a set of objects in a family tree (dendrogram) according to their similarity.

In order to work, an agglomerative hierarchical clustering algorithm needs to measure the similarity between objects that have to be clustered, similarity which in its turn is given by a distance function defined on the set of respective objects.

In our experiments we use rank distance (Dinu, 2003) to reflect stylistic similarity between texts. As style markers it use the function word frequencies. Function words 2.1. are generally considered good indicators of style because their use is very unlikely to be under the conscious control of the author and because of their psychological and cognitive role (Chung and Pennebaker, 2007). Also function words prove to be very effective in many author attribution studies. The novelty of the distance measure resides in the way it use the information given by the function word frequencies. Given a fixed set of function words (usually the most frequent ones), a ranking of these function words according to their frequencies is built for each text; the obtained ranked lists are subsequently used to compute the distance between two texts. To calculate the distance between two rankings we used *Rank distance* (Dinu, 2003), an ordinal distance tightly related to the so-called *Spearman's footrule*.

Usage of the ranking of function words in the calculation of the distance instead of the actual values of the frequencies may seem as a loss of information, but we consider that the process of ranking makes the distance measure more robust acting as a filter, eliminating the *noise* contained in the values of the frequencies. The fact that a specific function word has the rank 2 (is the second most frequent word) in one text and has the rank 4 (is the fourth most frequent word) in another text can be more relevant than the fact that the respective word appears 349 times in the first text and only 299 times in the second.

Rank distance (Dinu, 2003) is an ordinal metric able to compare different rankings of a set of objects.

A ranking of a set of n objects can be represented as a permutation of the integers $1, 2, \ldots, n$, $\sigma \in S_n$. $\sigma(i)$ will represent the place (rank) of the object i in the ranking. The Rank distance in this case is Spearman footrule:

$$D(\sigma_1, \sigma_2) = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)| \qquad (1)$$

This is a distance between what is called full rankings. However, in real situations, the problem of *tying* arises, when two or more objects claim the same rank (are ranked equally). For example, two or more function words can have the same frequency in a text and any ordering of them would be arbitrary.

The Rank distance allocates to tied objects a number which is the average of the ranks the tied objects share. For instance, if two objects claim the rank 2, then they will share the ranks 2 and 3 and both will receive the rank number $(2 + 3)/2 = 2.5$. In general, if k objects will claim the same rank and the first x ranks are already used by other objects, then they will share the ranks $x + 1, x + 2, \ldots, x+k$ and all of them will receive as rank the number: $\frac{(x+1)+(x+2)+\ldots+(x+k)}{k} = x + \frac{k+1}{2}$. In this case, a ranking will be no longer a permutation ($\sigma(i)$ can be a non integer value), but the formula (1) will remain a distance (Dinu, 2003).

Rank distance can be used as a stylistic distance between texts in the following way:

First a set of function word must be fixed. The most frequent function words may be selected or other criteria may be used for selection. In all our experiments we used a set of 120 most frequent Romanian function words.

Once the set of function words is established, for each text a ranking of these function word is computed. The ranking is done according to the function word frequencies in the text. Rank 1 will be assigned to the most frequent function word, rank 2 will be assigned to the second most frequent function word, and so on. The ties are resolved as we discussed above. If some function words from the set don't appear in the text, they will share the last places (ranks) of the ranking.

The distance between two texts will be the Rank distance between the two rankings of the function words corresponding to the respective texts.

Having the distance measure, the clustering algorithm initially assigns each object to its own cluster and then repeatedly merges pairs of clusters until the whole tree is formed. At each step the pair of nearest clusters is selected for merging. Various agglomerative hierarchical clustering algorithms differ in the way in which they measure the distance between clusters. Note that although a distance function between objects exists, the distance measure between clusters (set of objects) remains to be defined. In our experiments we used the *complete linkage* distance between clusters, the maximum of the distances between all pairs of objects drawn from the two clusters (one object from the first cluster, the other from the second).

2.3. New Testament clustering

We used the clustering with Rank distance to cluster the New Testament books. The resulted dendrogram is shown in Figure 2.3..

2.4. Pauline epistles and Petrine epistles

We used also the clustering with Rank distance to cluster the Paul's epistles. The resulted dendrogram is shown in Figure 2.4..

Several observations are immediate. First it stands out Philemon as forming an individual cluster. This can be explained by the fact that this is the only letter which is sure that was written by Paul (he wrote Philemon while being in jail, and therefore this was not dictated to his disciples and they could not have acted on it.

Then we note the first cluster which consists of II Timothy, Titus and I Timothy. The fact that Titus and 1 Timothy pattern alike strengthens the hypothesis that these were written by the same author.

The next cluster corresponds to I and II Thessalonians.

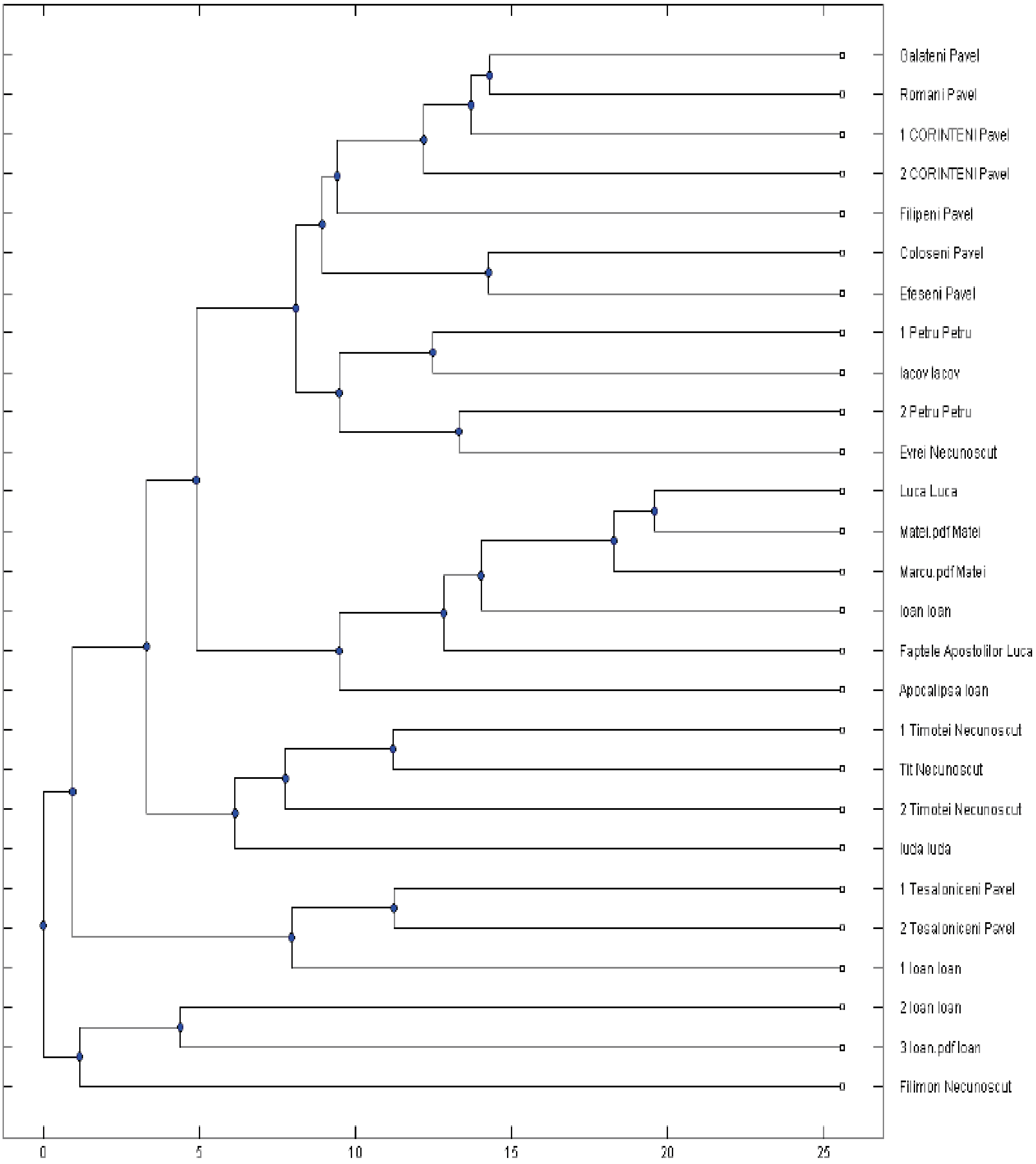


Figure 2: Dendrogram of New Testament books

The next five texts (Philippians, I and II Corinthians, Galatians, Romans) form a group that corresponds to the five grand letters. The last cluster includes two groups, first group corresponds to Hebrew and II Peter, and Peter's last epistle match Ephesians and Colossians (the latter being grouped together).

3. Conclusions and future works

In this paper we used two strategies in order to investigate the authorship identification in the Apostles' Epistles. We used a Romanian translation of the New Testament (Cornilescu, 1921) and applied the PCA and clustering analysis on these works. The results are consistent with some theological theories, bringing a plus of quantification and rigor. A critique of present research may be the fact that we used a Romanian translation of New Testament. In future works we intend to extend our research to the original (Greek) source texts. We want also to use other translations and to compare the obtained results to see how much the translation can influence the results. We want to investigate and to develop other different distances and techniques which are independent on the language.

Acknowledgments. The authors want to thank to anonymous reviewer for his helpful comments. We also thank to Cristian Panait for creating a part of the software application and to Maria-Octavia Şulea for her helpful comments. All authors contributed equally to this work. The research

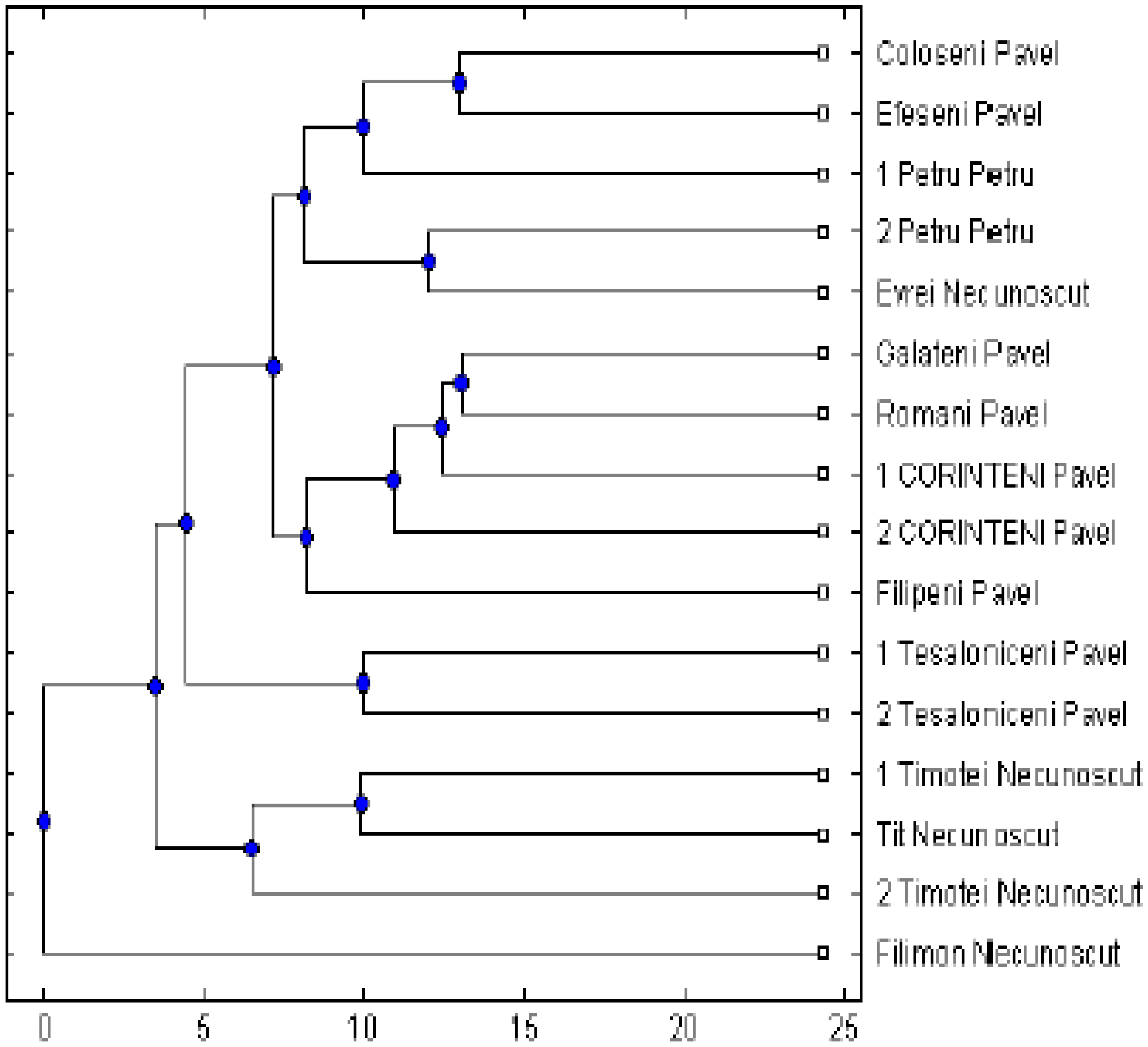


Figure 3: Dendrogram of Paul and Peter epistles

was supported by the CNCS, IDEI - PCE project 311/2011, ”The Structure and Interpretation of the Romanian Nominal Phrase in Discourse Representation Theory: the Determiners.”

4. References

G.K. Barr. 2003. Two styles in the new testament epistles. *Lit Linguist Computing*, 18(3):235–248.

F.F. Bruce. 1964. *The epistle to the Hebrews*. William B. Eerdmans Publishing Company.

C. K. Chung and J. W. Pennebaker, 2007. *The psychological function of function words*, pages 343–359. Psychology Press, New York.

D. Cornilescu. 1921. *Romanian Bible*. CSLI Publications.

A. Dinu and L. P. Dinu. 2005. On the syllabic similarities of romance languages. In *CICLing*, pages 785–788.

L.P. Dinu, M. Popescu, and A. Dinu. 2008. Authorship identification of romanian texts with controversial paternity. In *LREC*.

L.P. Dinu. 2003. On the classification and aggregation of hierarchies with different constitutive elements. *Fundam. Inform.*, 55(1):39–50.

R.O. Duda, P.E. Hart, and D.G. Stork. 2001. *Pattern Classification*. Wiley-Interscience Publication.

R. G. Gregson. 1966. A solutions to the problems of the thessalonian epistole. *Evangelical Quarterly*, 38:76–93.

D. Guthrie. 1961. *New Testament Introduction:the Paulin Epistles*.

J. Herring. 1970. *The epistle to the Hebrews*.

Hewitt. 1960. *The epistle to the Hebrews*. TNTC.

J. Bartont (eds.) J. Muddiman. 2010. *The Pauline Epistles*. Oxford Univ. Press.

Matthew L. Jockers, Daniela M. Witten, and Craig S. Criddle. 2008. Reassessing authorship of the book of mormon using delta and nearest shrunken centroid classification. *Lit Linguist Computing*, 23:465–491.

J. Knox. 1959. *Philemon among the Letters of Paul*.

M. Koppel, J. Schler, and E. Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8:1261–1276.

C. Labbe and D. Labbe. 2006. A tool for literary studies: Intertextual distance and tree classification. *Literary and Linguistic Computing*, 21(3):311–326.

T.W. Manson. 1962. *Studies in the Gospels and Epistles*.

C. L. Milton. 1955. *The Formation of the Pauline Corpus of Letters*.

F. Mosteller and D.L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison- Wesleys, Standford.

C. F. D. Moule. 1971. *La Genese du Nouveau Testament*. Delachaux & Niestle Editeurs.

M. Popescu and L. P. Dinu. 2008. Rank distance as a stylistic similarity. In *COLING (Posters)*, pages 91–94.

Behnam Sadeghi. 2011. The chronology of the qurn: A stylometric research program. *Arabica*, 210-299:465–491.

Hans van Halteren, R. Harald Baayen, Fiona J. Tweedie, Marco Haverkort, and Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77.

A. von Harnack. 2007. *Originea Noului Testament (The Origin of the New Testament)*. Herald, Bucharest.

R. McL. Wilson. 1958. *The Gnostic Problem*.

Automatic classification of Islamic Jurisprudence Categories

[1]Mohammad Hossein Elahimanesh, [2]Behrouz Minaei-Bidgoli, [3]Hossein Malekinezhad

[1] Islamic Azad University, Qazvin Branch, Qazvin, Iran

[2] Iran University of Science and Technology, Tehran, Iran

[3] Islamic Azad University, Naragh Branch, Naragh, Iran

[1,2,3]Computer Research Center of Islamic Sciences, Qom, Iran

E-mail: {elahimanesh, bminaei, hmalekinejad}@noornet.net

Abstract

This paper evaluates some of text classification methods to classify Islamic jurisprudence classes. One of prominent Islamic sciences is jurisprudence, which explores the religious rules from religious texts. For this study the Islamic Jurisprudence corpus is used. This corpus consists of more than 17000 text documents covering 57 different categories. The major purpose of this paper is evaluating text to numerical vectors converting methods and evaluating different methods of calculating proximity matrix between text documents for religious text classification. The results indicate that the best classification efficacy is achieved especially when 3-grams indexing method and KNN classifier using cosine similarity measure are applied. We reported 87.3% performance for Islamic jurisprudence categories classification.

Keywords: Text Classification, K-Nearest neighbor, Islamic Jurisprudence, N-grams

1. Introduction

One of the common processes in the field of text mining is text classification. As a simple definition, text classification detects the class of a new text based on a sufficient history of tagged and classified texts. Religious and historical text classification is one of important applications of text classification systems. The presence of religions in human societies has been caused the emergence of new branches of science during the different periods of time. The scientific results of these new scientific branches are mixture of historical and religious texts. Due to the huge volume of these text datasets, we should use of intelligent text processing techniques to explore the useful information with minimum time and space expenses. The rest of this paper is organized as follows. Related work is briefly reviewed in section 2. In section 3, The Schema tree of jurisprudence which used in our experiments is introduced. Our approach for text classification is detailed in section 4. Section 5 describes some conducted experiments to demonstrate the suitability of the proposed approach and finally section 6 concludes this paper.

2. Related Works

Different algorithms have been used in previous researches for religious text classification. Some of them are: Bayesian classifier, Support Vector Machine classifier, and K-Nearest Neighbor and Neural Network classifiers. Harrag et.al, have used a type of neural networks for Arabic texts classification (2009). Bayesian and Support Vector Machine classifiers have been used in Alsaleem researches (2011).

We can divide previous Islamic religious text classifiers into two fields: Quran and Hadith. Stem extended classifier is a sample of classifiers which is used for Hadith texts classification (Jbara, 2011). This classifier has been applied to classify Sahih AL-Bukhari Hadith book for 13 different categories. The performance of classifier is reported about 60%.

Another study on Sahih AL-Bukhari's Hadith book is done by Al-Kabi and his colleagues. They used of TF-IDF weighting method and reported 83.2% average accuracy for 8 different categories classification (2005). Al-Kabi and Al-Sinjilawi in another research on Sahih AL-Bukhari's Hadith book, have reported 85% classification accuracy for 12 different categories classification (2007).

Prophet Mohammad' Scripture is Quran. The context of Quran is divided to 30 parts, 114 chapters and 6236 verses. We can divide classifiers that have been applied on Quran into 4 groups based on the type of classes that they have identified. The classifiers of first group classify the chapters into two classes: Makki and Madani. Makki chapters in Mecca and Madani chapters in Medina have been revealed on the prophet Mohammad. Nassourou has done this type of classification on Quran (2011). The classifiers of second group classify the verses of Quran based on 15 different subjects of Islamic sciences. Al-Kabi applied this type of classification on the verses of two chapters of Quran. The accuracy of classification that he has reported is about 91% (2005). Another two groups of classifiers that have been studied by Nassourou are those classify the verses of Quran based on the dignity of descent and those classify the chapters based on the date of descent (2011).

3. Schema Tree of Jurisprudence

One of the prominent Islamic sciences is jurisprudence, which explores the religious rules from religious texts. In most cases each book has a table of contents it can be showed as a tree based on the subjects of book and it is possible to produce a type of subjective classification. Nevertheless this is so scarce that could be find two books with the same schema trees. Hence, one of activities of the computer research center of Islamic sciences is to produce the books independent schema trees using the researcher of each subject field. In this schema trees, only the headings that are in direct contact with desired knowledge have been mentioned.

One of these schema trees that are developed quickly is schema tree of jurisprudence. This schema tree has 50 main branches in various fields such as judicial orders, trading orders, food orders, family orders and so on. Each of these main branches includes many sub branches and leafs. The mentioned schema tree is connected to 9 books from important jurisprudence references so far and efforts to connect it to other references are attempting. These nine books are part of the Muslim hadith books.

4. Proposed method

The method that is used for converting documents to numerical vectors and the type of classifier that is applied, are two affecting factors in text classification systems. Proposed method in this paper investigates the effects of two ways of converting documents to numerical vectors on text classification performance. Also the effects of two similarity measures, Cosine and Dice measures, in combination with KNN classifier have been studied.

4.1 Converting Text Documents into Numerical Vectors

In this paper, we used N-grams of characters and N-grams of words to convert text documents into numerical vectors. Using N-grams consist of characters is a common way to represent text documents as numerical vectors. In this technique, each text document divides into slices with length N of adjacent characters. Vector corresponding to each document contains a list of non-iterative N-Grams with the number of iterations of each N-gram. In previous studies, the common lengths of N-grams were between 2 and 5 but in this paper we investigate the effect of N-grams with various lengths, between 2 and 10, on the efficacy of classifier.

Word N-grams of a text are the N-word slices of text that are generated from the adjacent words. Each word N-gram is a dimension in equivalent numerical vector of document. Corresponding value of each dimension for a document is the number of occurrence of N-grams in that document. Word N-grams with lengths more than one can be considered as multi words. Multi words are used in some previous works. For example, Zhang used of multi words that was extracted using linguistic rules, to improve the English text classification (2008). In this study we replace the Corresponding value of each dimension in equivalent numerical vector of document with the TF-IDF weight corresponding with each dimension.

4.1.1. TF-IDF Weighting

TF-IDF weighting is a common weighting method in the field of text mining. This weighting method calculates the importance of N-grams of a document than the corpus is used. The N-grams with greater TF-IDF are more important. The exact formulation for TF-IDF is given below

$$\text{TF-IDF (N-gram)} = \text{tf(N-gram)*idf(N-gram)}$$

Where tf(N-gram) is the number of occurrences of n-gram in document, and idf(N-gram) calculates from

$$\text{Idf (N-gram)} = \log \quad |D| / |\{d : \text{N-gram} \in d\}|$$

Where |D| is the number of documents in training dataset and | {d : N-gram ∈ d}| is the numbers of documents contain the N-gram.

4.2 KNN Classifier

K-Nearest neighbor (KNN) classification finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood. Given a training set D and a test object X, the algorithm computes the similarity between X and all the training objects to determine its nearest-neighbor list. It then assigns a class to X by taking the class of the majority of neighboring objects. We use following formulas to calculate probability of sample X belong to each class

$$\mathrm{P}(\mathrm{X}, C_j) = \frac{1}{K} . \sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{SIM}(\mathrm{X}, \mathrm{d_i}) . \mathrm{y}(\mathrm{d_i}, \mathrm{C_j})$$

Where $\mathrm{d_i}$ is the $\mathrm{i^{th}}$ document of training documents dataset and $\mathrm{y}(\mathrm{d_i}, \mathrm{C_j})$, Represents the belonging of document $\mathrm{d_i}$ to category $\mathrm{C_j}$ and calculate from the following formula

$$\mathrm{y}(\mathrm{d_i}, \mathrm{C_j}) = \begin{cases} 1, \mathrm{d_i} \in \mathrm{C_j} \\ 0, \text{Otherwise} \end{cases}$$

We calculate the similarity between document d_i and test sample X, SIM(X, $\mathrm{d_i}$), using two similarity measures, Dice and Cosine, as following

$$\mathrm{Dice}(\mathrm{X}, \mathrm{d_i}) = \frac{2 \times |\mathrm{X} \cap \mathrm{d_i}|}{|\mathrm{X}| + |\mathrm{d_i}|}$$

$$\mathrm{Cosine}(\mathrm{X}, \mathrm{d_i}) = \frac{\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{X_j} \times \mathrm{d_{ij}}}{\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{n}} (\mathrm{X_j})^2} \times \sqrt{\sum_{\mathrm{j}=1}^{\mathrm{n}} (\mathrm{d_{ij}})^2}}$$

Where $|\mathrm{X} \cap \mathrm{d_i}|$ is the number of common N-grams between test sample X and training sample d_i. In Cosine measure the number of training documents is n. X_j *and* d_{ij} respectively are TF-IDF weights corresponding to $\mathrm{j^{th}}$ dimension of test sample X and TF-IDF weights corresponding to training document d_i. Finally, Sample X belongs to the class that has the largest $P(X, C_j)$.

5. Implementation and Results

A variety of experiments were conducted to test the performance of proposed method. Accuracy of the classifier is measured by a 5-fold cross-validation procedure. Essentially, the corpus is divided into 5 mutually exclusive partitions and the algorithm is run once for each partition. Each time a different partition is used as the test set and the other 4 partitions are used as the training set. The results of the 5 runs are then averaged. We first explain the training and test data used in the experiments and then present the obtained results.

5.1 Islamic Jurisprudence corpus

The Islamic Jurisprudence corpus consists of more than 170000 text documents covering 57 different categories. This corpus developed by Islamic researcher in CRCIS to the next research of any Islamic researchers will be easier. Document in this corpus contains one paragraph in average (Less than 250 words).

We choose 11699 text documents covering 9 categories (this number of document is total available document that can be used). The detailed specification of the corpus is shown in Table1. As the Table 1 indicates, the corpus has the imbalanced distribution of documents in categories. We will show the effect of imbalanced categories on classification performance for each category. Average length of documents is equal to 137 characters and average length of each word is equal to 4 characters and approximately each document consists of 27 words.

Class Number	Class Name	Number of Documents
1	Retribution	1017
2	Hajj	2407
3	Jobs	547
4	Heritage	779
5	Marriage	996
6	Praying	2429
7	Fasting	699
8	Cleanliness	2306
9	Almsgiving	519

Table 1: Detailed description of the corpus

As the Table 1 indicates, the corpus has the imbalanced distribution of documents in categories. We will show the effect of imbalanced categories on classification performance for each category. Average length of documents is equal to 137 characters and average length of each word is equal to 4 characters and approximately each document consists of 27 words.

5.2 Evaluation Measures

In the text classification, the most commonly used performance measures are precision, recall and F-measure. Precision on a category is the number of correct assignments to this category and recall on a category signifies the rate of correct classified documents to this category among the total number of documents belonging to this category. There is a trade-off between precision and recall of a system. The F-measure is the harmonic mean of precision and recall and takes into account effects of both precision and recall measures. To evaluate the overall performance over the different categories, micro and macro averaging can be used. In macro averaging the average of precision or recall is compared over all categories. Macro averaging gives the same importance to all the categories. On the other hand micro averaging considers the number of documents in each category and compute the average in proportion to these numbers. It gives the same importance to all the documents. When the corpus has unbalanced distribution of documents into categories, by using macro averaging, classifier deficiency in classifying a category with fewer documents is emphasized. Since an imbalanced corpus is being dealt with, it seems more reasonable to use micro averaging.

5.3 Experimental Results

In this section, the experimental results are presented. The experiments consist of evaluating classifier performance when word n-grams and character n-grams are used to represent documents. The effects of using cosine and dice similarity measures have been analysed also.

In first experiment we have used of word n-grams with various lengths in preprocessing and then we have applied KNN classifier on the data. Fig.1 shows the best performance that can be achieved is in the case that we choose word N-grams with length 1 and is equal to 81.6%.

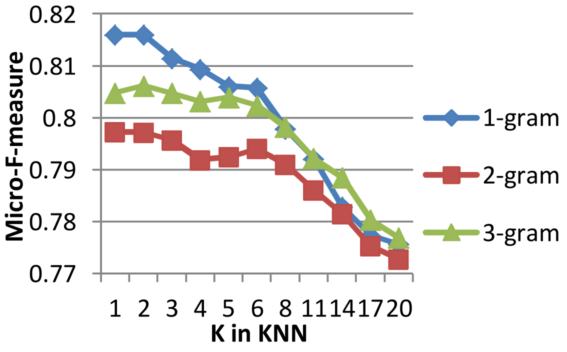


Figure 1: Evaluation of classifier with different word N-grams lengths

In second experiment we have used of character N-grams in preprocessing and then we have applied KNN classifier on the data. The results have been shown in Fig.2. As the results indicate, best performance of classifier would happen when we use 3-grams in preprocessing stage.

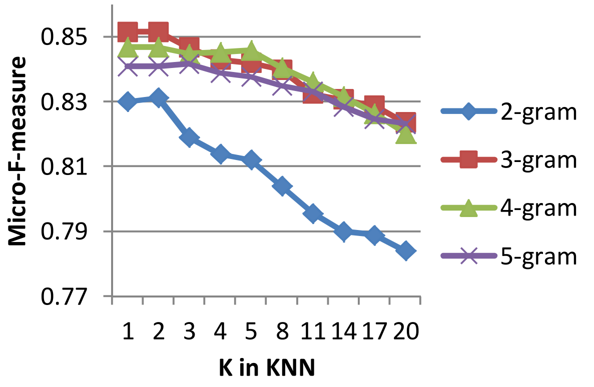


Figure 2: Evaluation of classifier with different character N-grams lengths

The effects of using different similarity measures, cosine and dice measures, on the classifier performance are evaluated in last experiment. In this experiment we have used of character 3-grams in preprocessing, because the results of second experiment showed character N-grams provide better performance in comparison with word N-grams. The results have been shown in Fig.3.

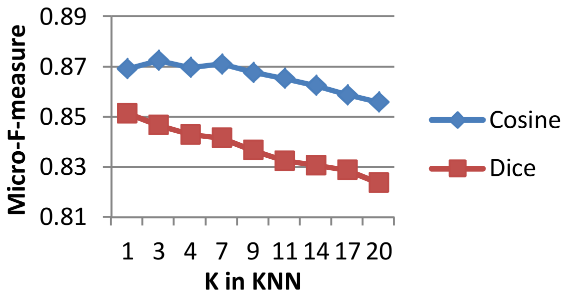


Figure 3: Evaluation of classifier with different similarity Measures

Table 2 is a comparison between obtained classification results for each category of datasets when applying classifier using character 3-grams and cosine similarity measure.

Some of different categories of jurisprudence are so similar and this cause to produce incorrect classification results. Confusion matrix of proposed method is shown in Table 3.

Class Name	Precision	Recall	Accuracy	F-Measure
Retribution	90.9%	91.8%	98.5%	91.4%
Hajj	88.4%	91.6%	95.8%	90%
Jobs	86.7%	81.7%	98.6%	84.1%
Heritage	87.5%	91.1%	98.5%	89.3%
Marriage	83.9%	82.9%	97.2%	83.4%
Praying	87.7%	85.9%	94.6%	86.8%
Fasting	80.2%	77.8%	97.5%	79%
Cleanliness	85.9%	83.1%	98.6%	84.4%
Almsgiving	85.9%	83.1%	98.6%	84.4%

Table 2: Comparison between obtained classification results for each category

		Predicted Class									
		Almsgiving	Cleanliness	Fasting	Praying	Marriage	Heritage	Jobs	Hajj	Retribution	Total
Actual Class	Retribution	1	3	2	2	3	4	2	3	187	207
	Hajj	3	10	4	12	5	2	2	441	4	483
	Jobs	2	3	2	3	7	1	89	3	2	112
	Heritage	2	2	0	2	4	142	1	1	4	158
	Marriage	2	6	2	3	165	8	4	6	3	199
	Praying	3	24	11	417	4	2	2	19	3	485
	Fasting	1	6	109	13	3	1	2	6	1	142
	Cleanliness	2	405	5	22	5	3	2	17	3	464
	Almsgiving	86	2	3	4	2	3	2	4	2	108
Total		102	461	138	478	198	166	106	500	209	2358

Table3: Confusion matrix of proposed method

6. Conclusion

The proposed approach in this paper aims to enhance the classification performance of KNN classifier for religious text classification. As the results indicate, this approach improves the text classification especially when character 3-grams indexing method and cosine similarity measure are applied. We reported 87.25% performance for Islamic Jurisprudence corpus classification while the performance of previous reported study on Sahih AL-Bukhari's Hadis books that is so similar to our corpus, is equal to 60%.

7. Acknowledgements

The authors would like to thank Noor Text Mining Research group of Computer Research Center of Islamic Sciences for supporting this work.

8. References

Harrag, F., El-Qawasmah, E. (2009). Neural Network for Arabic Text Classification. *In Proceeding of ICADIWT '09,* pp. 778--783.

Alsaleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. *In Proceeding of 2'th IAJeT*, Vol. 2 ,June 2011.

Jbara, K. (2011). Knowledge Discovery in Al-Hadith Using Text Classification Algorithm. *Journal of American Science*.

Al-Kabi, M. N., Kanaan, G., Al-Shalabi, R. (2005). Al-Hadith Text Classifier. *In Proceeding of 5'th Journal of Applied Sciences*, pp. 584--587.

AL-Kabi, M. N., AL-Sinjilawi, S. I. (2007). A Comparative Study of the Efficiency of Different Measures to Classify Arabic Text. *University of Sharjah Journal of Pure and Applied Sciences*, 4(2), pp. 13--26.

Nassourou, M. (2011). Using Machine Learning Algorithms for Categorization Quranic Chapters by Major Phases of Prophet Mohammad's Messengership. *Published by University of Würzburg*, Germany.

Al-Kabi, M. N., Kanaan, G., Al-Shalabi, R. (2005). Statistical Classifier of the Holy Quran Verses (Fatiha and Yaseen Chapters). *In Proceeding of 5'th Journal of Applied Sciences*, pp. 580-583.

Nassourou, M. (2011). A Knowledge-based Hybrid Statistical Classifier for Reconstructing the Chronology of the Quran. *Published by University of Würzburg*, Germany.

Zhang, W., Yoshida, T., Tang, X. (2008). Text Classification based on multi-word with support vector machine. *Knowledge-Based Systems*, vol.21, pp. 879-886.

Correspondence Analysis of the New Testament

Harry Erwin, Michael Oakes
University of Sunderland
DCET, DGIC, St. Peter's Campus, St. Peter's Way, Sunderland SR6 0DD, England
E-mail: michael.oakes@sunderland.ac.uk

Abstract

In this paper we describe the multivariate statistical technique of correspondence analysis, and its use in the stylometric analysis of the New Testament. We confirm Mealand's finding that texts from Q are distinct from the remainder of Luke, and find that the first 12 chapters of Acts are more similar to each other than to either Luke or the rest of Acts. We describe initial work in showing that a possible "Signs Gospel", describing Jesus' seven public miracles, is indeed distinct from the remainder of John's Gospel, but that the differences are slight and possibly due to differences in genre.

1. Introduction

Correspondence analysis is a multivariate statistical technique, originally developed by Benzécri (1980). When it is used for the comparison of texts, we start with matrix of whole numbers where the rows correspond to the text samples, and each column corresponds to a countable linguistic feature of that text. In the studies described in this paper, the text samples are 500-word samples of the New Testament in the original Greek, and the columns are word counts for each of the 75 most common words[1] in the Johannine corpus (John's Gospel and Epistles, Revelation). The technique of correspondence analysis takes into account the fact that many linguistic features vary together – texts containing many occurrences of one feature such as the word ημων also contain many occurrences of γαρ and υμων. By considering such similarly distributed groups of linguistic features as one, texts originally described by a large number of features can be plotted on a graph determined by just two factors or groups of co-occurring features. The individual words constituting each factor can also be plotted on the same graph, as shown in Figure 1. At the top of the figure, the text samples from Revelation (labelled "r") score most highly on the second factor (y axis), which s characterised by high occurrence of such words as επι, επτα and γησ. They also have small positive scores on the first factor (x axis), which is most characterised by such words as αυτον and ειπεν. Commands in the R statistical programming language for performing correspondence analysis on texts are given by Baayen (2008:128-136). We used the Westcott and Hort Greek texts[2], stripped out the apparatus and the verse numbers, and then broke each test up into sequential samples of a fixed length of 500 Greek words (except for the final sample from each book, which is shorter). Correspondence analysis cannot be performed if any of the values in the input matrix are 0, so we increased all the frequency counts in the entire matrix by 1.

As seen in Figure 1, our correspondence analysis for the entire New Testament grouped the text samples into four main clusters: Revelation ("r"), the Synoptic Gospels ("Mt", "Mk", "Lk"), the Epistles or letters ("l") and the Gospel of John ("jg" and "s", see section 5). These patterns suggest that the technique is trustworthy, since texts that we would expect to be similar do indeed group together, away from groups of dissimilar texts. Secondly, we see that genre plays a major role in grouping texts. For example, the letters all group together although written by various authors; similarly the Synoptic Gospels group together despite having been written by three different authors. Finally, we see that the texts of Revelation are quite distinct from those of John's Gospel, supporting the general opinion that it has a separate author. Having used correspondence analysis to gain an overview of the main text groupings in the New Testament, in the remainder of this paper we will look at some of the main questions of authorship that have been considered by New Testament scholars. In section 2 we will examine evidence for Q, a proposed source of the Synoptic Gospels. In section 3 we will attempt to answer the question "Did Luke write Acts?". In Section 4 we will discuss the extent of the Pauline corpus, and in Section 5 we will examine whether correspondence analysis throws any light on the question of whether the Gospel of John draws on an earlier source called the "Signs Gospel". We will conclude with some thoughts on how to control for genre effects which can swamp the effects of individual writing styles.

[1] www.mrl.nott.ac.uk/~axc/DReSS_Outputs/LFAS_2008.pdf

[2] http://www-user.uni-bremen.de/%7Ewie/GNT/books.html

2. Q

A widely, but not universally, held view is that the Gospels of Matthew and Luke each draw both upon the Gospel of Mark and a second, probably oral, Greek source. This second original source is referred to as Q, which stands for "quelle" or "source" in German. A detailed recent review of the computational evidence both for and against the existence of Q is given by Poirier (2008). The authors were first alerted to the potential of correspondence analysis to determine the provenance of texts by Mealand's (1995) use of this technique to show that material thought to be from Q is fairly distinct from both those sections from Luke which also appear in Mark (labelled "m"), and those sections which are unique to Luke (labelled "L"). For comparison, he also included samples of Mark's Gospel (labelled "M"). Mealand presents his raw data at the end of his paper, where frequency counts for 25 parts of speech, common words, or other numeric linguistic data, are given for each of the text samples of Luke. The texts are labelled according to whether they constitute infancy narrative, genealogy, are common to Mark, are unique to Luke, or thought to come from Q. Figure 2 shows our own correspondence analysis using Mealand's data for the frequencies of the three Greek words "kai", "nou" and "aut" in each text sample. We were also able to achieve the broad separation of the Q samples from the other texts as described by Mealand. As previously found, one "m" sample was found far from the others, at the extreme left of the diagram. This sample is Luke 21:1-34, Jesus' apocalyptic speech – on its own, due to its being the sole member of that genre in that data set (Linmans, 1998:4). It is also more difficult to discriminate between the material in Luke common to Mark from that unique to Luke. A second source of numeric linguistic data obtained for a correspondence analysis on the "problem of Q" is given by Linmans (1995), where each text has counts for the 23 most common words and 20 parts of speech, and the text samples are classified into one of four genres: narrative, dialogue, aphorisms and parables.

3. Luke and Acts

Traditionally scholars have considered the book of Acts to have been written in its entirety by the author of the Gospel of Luke, since Luke was a companion of Paul, both prefaces are dedicated to "Theophilus", and the preface of Acts refers to a former book by its author. The narrative also follows on smoothly from the end of Luke to the start of Acts. However, Greenwood's (1995) computer study suggested that only the early chapters of Acts resemble Luke stylistically, while the later chapters, describing Paul's missionary journeys, are stylistically distinct. His technique was to use the hierarchical clustering algorithm of Hartigan and Wong on the frequencies of the most common words in the whole data set for each chapter of Acts. To investigate this question for ourselves, we performed a correspondence analysis on our set of 500-word samples for both Luke and Acts, taking into account the frequencies of the 75 most common Greek words in the Johannine corpus. Our results are shown in Figure 3, where the text samples numbered 1 to 38 are from the Gospel of Luke, those numbered 39 to 44 are from the first 12 chapters of Acts, and the remainder are from the latter part of Acts. It can be seen that while Luke and the latter part of Acts are largely distinct from each other, the early chapters of Acts are in an intermediate position. They cluster closely together, so clearly have much stylistically in common with each other, but appear to be distinct from both the Gospel of Luke and the later chapters of Acts.

4. The Pauline Epistles and Hebrews

Some of the Epistles of Paul are more widely thought to be authentic than others. The general consensus is that the four so-called "Hauptbriefe" (Romans, 1 and 2 Corinthians, and Galatians) are certainly authentic. Most scholars also accept 1 Thessalonians, Philippians and Philemon. The most disputed letters are Colossians, Ephesians, and 2 Thessalonians. The Pastorals (1 and 2 Timothy and Titus) and Hebrews are considered least likely to have been written by Paul, and indeed, the authorship of Hebrews is a complete mystery. Since the Reformation, Hebrews has been generally considered not to have been written by Paul, partly because unlike in his other letters, Paul does not introduce himself at the start. However, there is some kind of link with Paul, as Hebrews mentions Timothy as the author's companion. Computer studies of the writing style of the Pauline Epistles have been performed by Neumann (1990:191). Using the Mahalanobis distance of various texts from each other, based on a matrix of texts and the counts of a large number of linguistic features, he concluded that the Pastoral Epistles were not written in the style typical of St. Paul's more accepted writings. Using another multivariate technique called discriminant analysis, he compared the distances of disputed texts from the "Pauline centroid", the main cluster of accepted texts. This technique was thus a form of outlier analysis, and Neumann concluded that there was "little reason on the basis of style to deny authenticity" to the disputed letters Ephesians, Colossians and 2 Thessalonians. Other multivariate techniques, namely principal component analysis and canonical discriminant analysis were performed by Ledger (1995). He found that 1 and 2 Corinthians, Galatians, Philemon, 2 Thessalonians and Romans seem to form a "core Pauline group", while Hebrews was a definite outlier. He also felt that the authorship of all the remaining letters was doubtful. Greenwood (1992), again using the hierarchical clustering technique of Hartigan and Wong, found distinct clusters corresponding to the Missionary, Captivity and Pastoral letters.

The results of our own correspondence analysis of the New Testament epistles are shown in Figure 4. As well as those letters traditionally attributed to Paul, for comparison we included the letters of John ("Jn1", "Jn2" and "Jn3"), James ("jam"), Jude ("jude") and Peter ("1Pet", "2Pet"). The first letter of John was most clearly distinct from all the other letters, with relatively high occurrences of "εσπν" and "οπ". The four Hauptbriefe, ("1cor", "2cor", "rom" and "gal", all in darker type) all group together on the left hand side of the graph, suggesting that they make a homogeneous group. The disputed Pauline Epistles are mainly found on the right hand side of the graph, with Ephesians ("eph") and Colossians ("col") the most distinct from the Hauptbriefe. Although Hebrews ("heb") is not thought to be written by Paul, the Hebrews samples form a close cluster on the borderline between Paul's Hauptbriefe and the disputed letters. Differences in the styles of Paul's letters might have arisen through dictation to various amanuenses. We do know that on at lest one occasion Paul used dictation, as Romans 16:22 contains the words "I Tertius, the writer of this letter"[3].

5. The Signs Gospel

The term "Signs Gospel" was first used by C.H. Dodd (1963) to refer to chapters 2 to 12 of the Gospel of John. He used this name because these chapters describe Jesus' seven public miracles, which were signs of his messianic identity (Thatcher, 2001). Later, this "Signs Gospel" was thought to also consist of a Passion narrative. There are four main theories regarding the use of early sources in John's Gospel. Firstly, we have the oral tradition theory of Dodd and Thatcher themselves, which is that many sayings of Jesus were drawn from an oral tradition, some of which was also used by the writers of the Synoptic Gospels. The written source theory is that John's Gospel was drawn from two written sources, a miracle source and a version of the Passion story which had been combined before the time of John. These postulated sources have since been lost. The third theory is the synoptic dependence theory, in which the Gospel of John was also based on written sources, most clearly the Synoptic Gospels. The problem with this theory is that the differences between John's Gospel and the other three are much greater than the similarities between them, but recently the Leuven school have come to believe that there are some key correspondences such as Luke 24:12 and John 20:3-30. Fourthly, the developmental theory is that the Gospel was based on repeated editing by a Johannine community (Thatcher, 1989). Felton and Thatcher (1990) performed a stylometric analysis where the t-test was used to compare texts thought to be from the Signs Gospel with those from the remainder of John, according to the frequencies of certain linguistic features such as the number of definite articles in each text block, verb-verb sequences, and the number of words containing from each of 1 to 10 characters. Their results were inconclusive.

We used a correspondence analysis to determine whether the text of the putative Signs Gospel differs stylometrically from the rest of the book of John. Once again we used the Westcott and Hort original Greek text, and this time followed the reconstruction of the Signs Gospel given by Fortna (2010:189). Some of our findings can be seen in Figure 1, the correspondence analysis of the New Testament as a whole. The 500-word samples from John's Gospel are labeled "s" for the Signs Gospel, and "jg" for the rest of John's Gospel except for the 3 samples labeled "jf" which come from a passage in John known as the "Farewell Discourse". For comparison, the letters of John are labeled "jl", in contrast to all the other letters in the New Testament which are simply labeled "l". Nearly all the samples from John's Gospel are low down in the south-east quadrant of the diagram, showing that the writing style in John's Gospel as a whole is quite distinct from the rest of the New Testament. The "s" samples are grouped close to each other, and so although it has been suggested that the Signs Gospel was originally composed by combining a life narrative with a passion narrative, there is no evidence for more than one author. The "s" samples are not far from the "jg" samples, but a t-test for matched pairs showed that there were significant differences between the sets of co-ordinates for both factor 1 ($p = 0.0005$) and factor 2 ($p = 0.0013$) for the two sets of samples. It remains for us to determine whether these small differences really

[3] http://ww2.ferrum.edu/dhowell/rel113/pauls_letters/pathway.htm

were due to distinct writing styles in the two sources, or were in some measure due to genre. Three samples from the Gospel of John did stand out, namely the “jf” samples of the “Farewell Discourse” (John 13:31 to 17:26). These samples had more in common stylometrically with the letters of John (which are widely thought to be authentically by John) than with the other parts of John’s Gospel.

6. Conclusions

In this paper we have described the use of correspondence analysis to first map out the main stylometric groupings (letters, Synoptic Gospels, Revelation and John) among the New Testament texts as a whole. We have shown that John in both his Gospel and letters is distinct in his use of common words from the other books of the New Testament. Revelation and the John samples are at opposite poles of the correspondence analysis plot, showing that as is commonly supposed, they authors are unlikely to be one and the same. We then examined specific controversies about New Testament authorship. We have reproduced Mealand’s finding that Q is stylometrically distinct from the rest of Luke, and shown that the first 12 chapters of Acts form a homogeneous group, intermediate in style between Luke and the rest of Acts. In our on-going study of the Signs Gospel, a possible precursor of John, we found that the Signs samples clustered very close together, suggesting single authorship of those samples. However, the positions on the plot of the Signs samples were only slightly different from the rest of John, and we have not yet controlled for the fact that these differences in position might be due to differences in genre. There are at least three suggestions in the literature for factoring out genre. One, by Mealand (2011), is that the first factor in a correspondence analysis, accounting for most of the variation between text samples, might be the one most due to genre differences, and thus later factors might be less affected. A plot where the axes are the second and third factors might then show more clearly differences in individual style than the more commonly displayed plots where the axes are the first and second factors. A second suggestion is simply to compare “like with like”. Thus for example, only samples in the same genre, such as narrative texts, should be compared in the same analysis. Thirdly, Linmans (1995, 1998) has proposed using correspondence analysis in conjunction with another multivariate technique, log linear analysis (LLA), to factor out genre differences.

The work described here stands in contrast with the work of Jockers et al. (2008) and Sadeghi (2011), who performed authorship studies on the Book of Mormon and the Quran respectively. These differed from our starting point in that the accepted assumption is that these texts have a single author, and stylometrics were used to help test this assumption.

7. References

Baayen, R. H. (2008). *Analysing Lingusitic Data*. 2008. Cambridge University Press.

Benzecri, J-P. (1980). *L’analyse des données*. Tome 2: L’analyse des correspondances. Paris: Bordas.

Felton, T. and Thatcher, T. (2001). Stylometry and the Signs Gospel. In R. T. Fortna and T. Thatcher (editors), *Jesus in Johannine Tradition*. Louisville-London: Westminster John Knox Press, pp. 209-218.

Fortna, R. T. (2010). The Signs Gospel. In: Robert J. Miller (editor), *The Complete Gospels*, Fourth Edition, Polebridge Press, Salem, Oregon.

Greenwood, H. H. (1995). Common word frequencies and authorship in Luke’s Gospel and Acts. *Literary and Linguistic Computing,* 10(3), pp. 183-187.

Greenwood, H.H. (1992). St. Paul revisited – A computational result, *Literary and Linguistic Computing*, 7(1), pp 43-47

Jockers, M.L., Witten, D.M. and Criddle, C.S. (2008). Reassessing Authorship of the Book of Mormon using Delta and Nearest Shrunken Centroid clustering. *Literary and Linguistic Computing* 23(4), pp. 465-492.

Ledger, G. (1995). An exploration of differences in the Pauline Epistles using multivariate statistical analysis. *Literary and Linguistic Computing* 10(2), pp. 85-96.

Linmans, A.J.M. (1995). *Onderschikking in de Synoptische Eevangelien.* Nederlandse organisatie voor wetenschappelijk onderzoek.

Linmans, A. J. M. (1998). Correspondence analysis of the Synoptic Gospels. *Literary and Linguistic Computing*, Vol. 13(1), pp. 1-13.

Mealand, D.L. (1995). Correspondence analysis of Luke, *Literary and Linguistic Computing* 10(3), pp. 171-182.

Mealand, D.L. (2011). Is there stylometric evidence for Q? *New Testament Studies*, 57, pp. 483-507.

Poirier. J. C. (2008). Statistical studies of the verbal agreements. *Currents in Biblical Research* 7(1), pp. 68-123.

Sadeghi, B. (2011). The chronology of the Quran: A stylometric research program. *Arabica* 58 (3-4), pp. 210-299.

Thatcher, T. (2001) Introduction to *Jesus in Johannine Tradition*, In Robert T. Fortna and Tom Thatcher (eds), Louisville and London: Westminster John Knox Press.

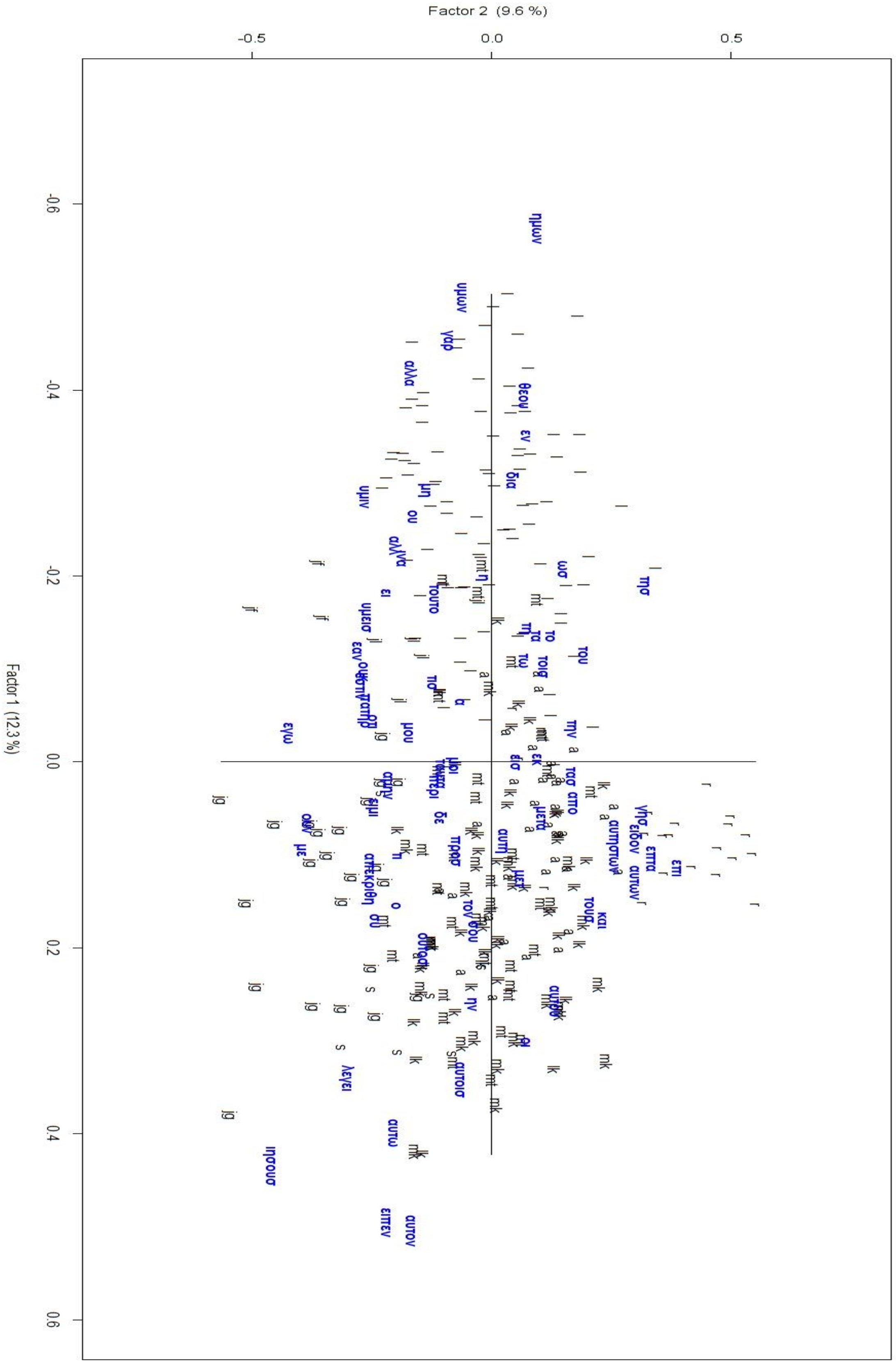


Figure 1. Overview of the New Testament by Correspondence Analysis. The four main clusters are Revelation ("r"), the Synoptic Gospels ("mt", "mk", "lk"), Epistles or Letters ("l") and the Gospel of John ("jg").

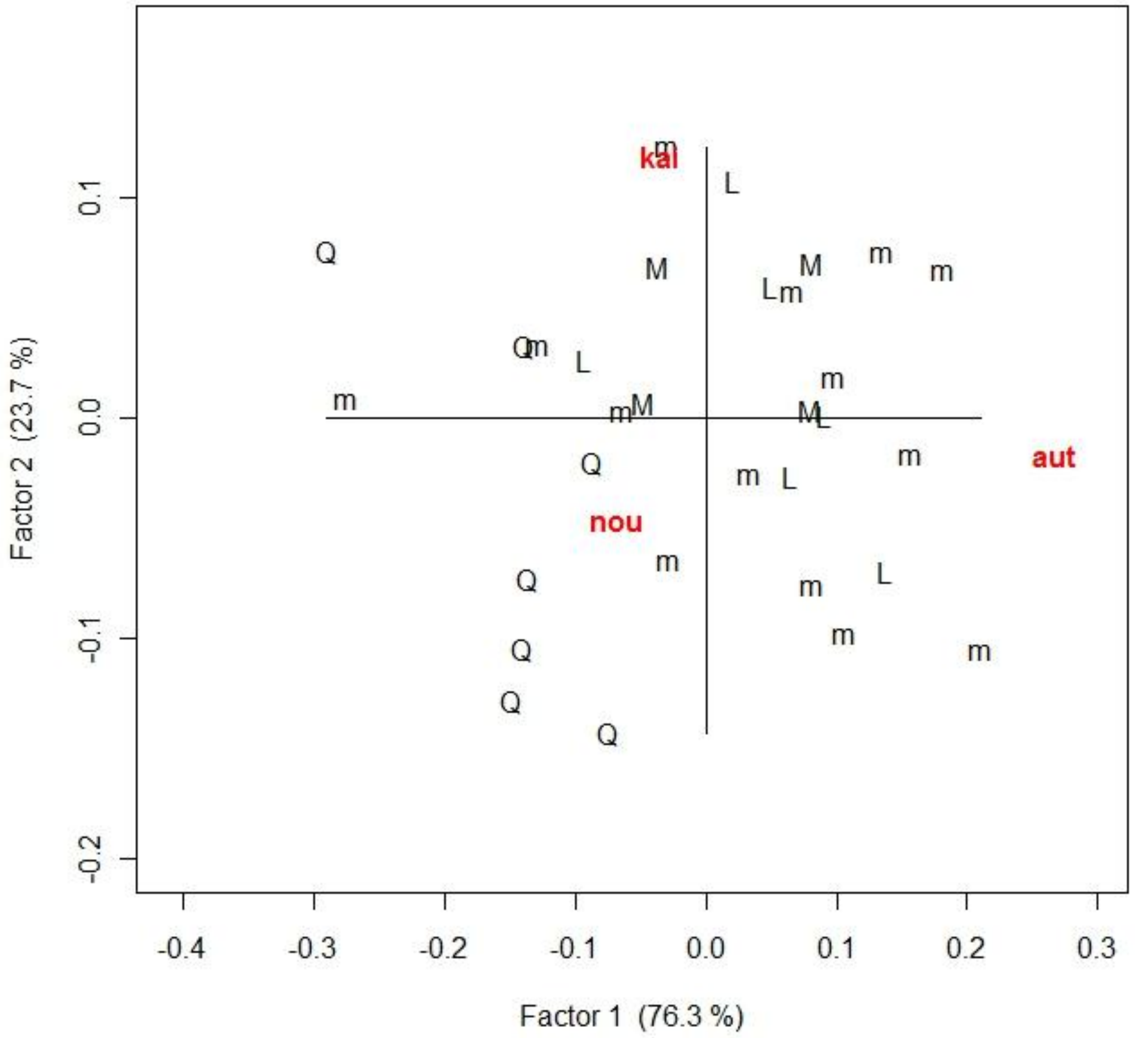


Figure 2. Evidence for Q from Mealand’s (1995) Data. The samples of Q broadly stand out from the other material in Mark and Luke.

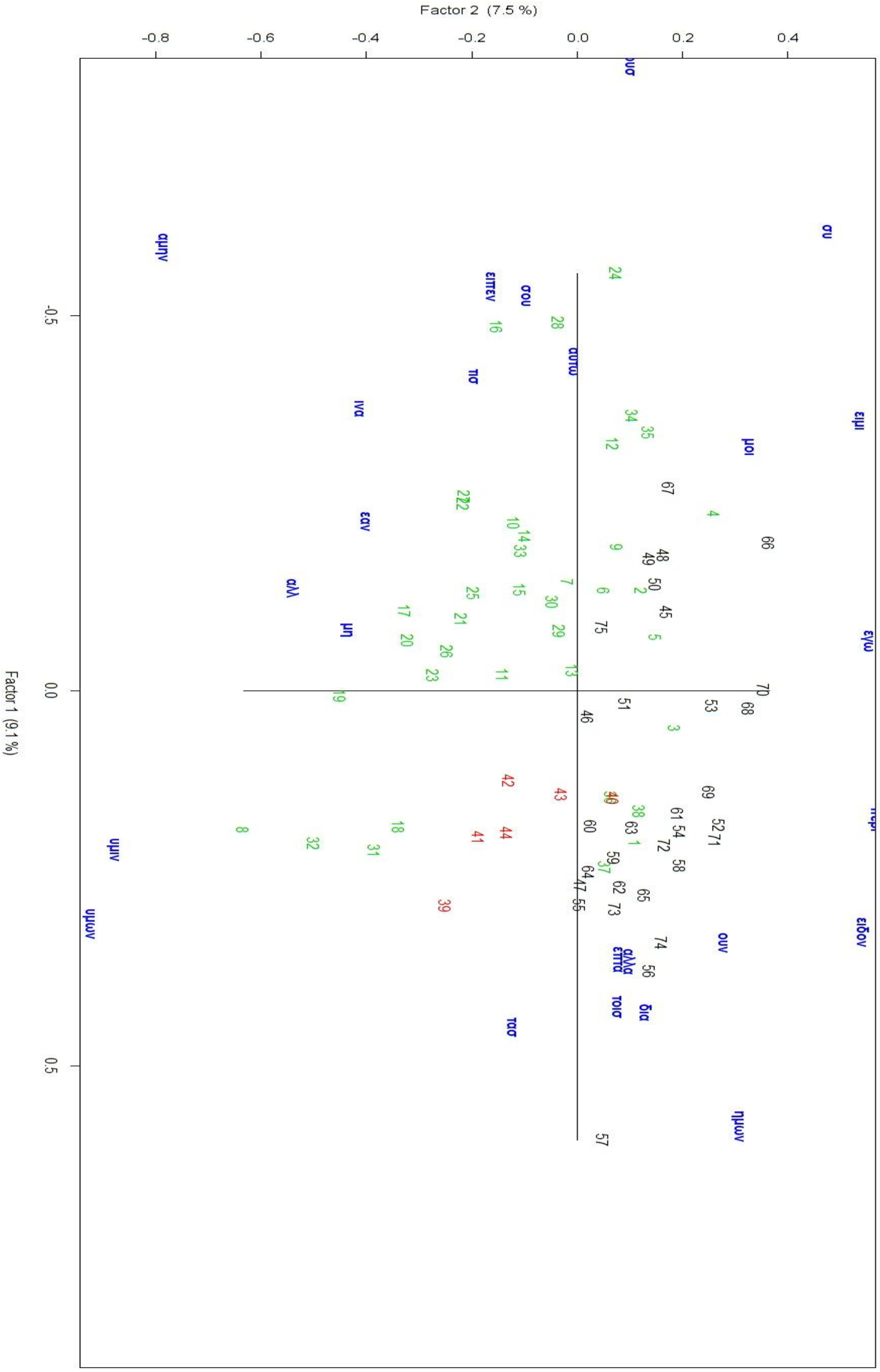


Figure 3. Comparison of Luke and Acts by Correspondence Analysis. Three main clusters are seen: Luke (samples 1-38), the first 12 chapters of Acts (samples 39-44), and the remainder of Acts (samples 45-75).

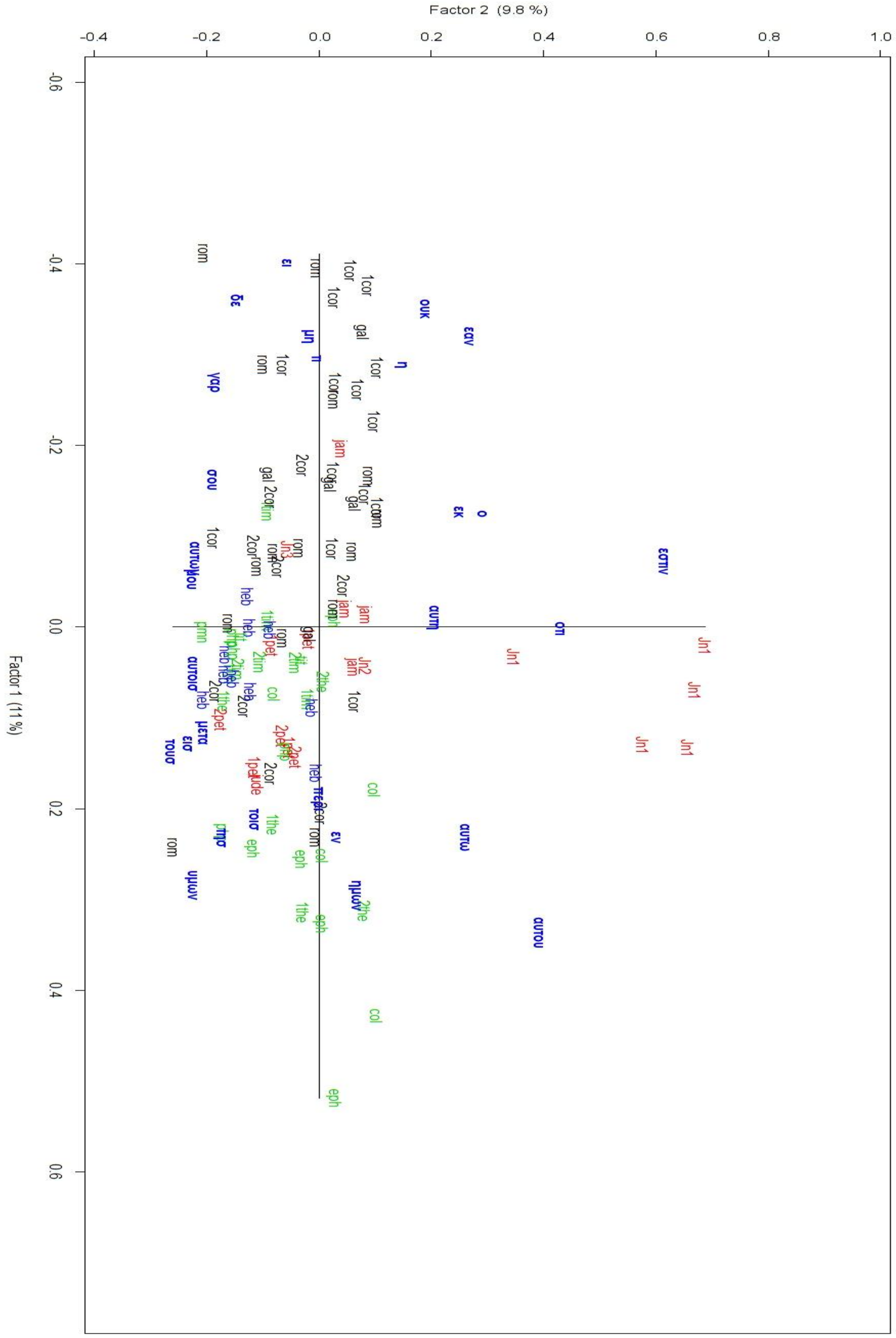


Figure 4. Comparison of the New Testament Epistles by Correspondence Analysis. The four *Hauptbriefe* ("1 cor", "2 cor", "rom", "gal") all group together on the left hand side of the graph, while Ephesians ("eph") and Colossians ("col") are the most distinct from the core Pauline group.

A new framework for detecting similar texts in Islamic Hadith Corpora

Hossein Juzi[1], Ahmed Rabiei Zadeh[2], Ehsan Barati[3], Behrouz Minaei-Bidgoli[4]
[1,2,3,4] Computer Research Center of Islamic Sciences, Qom, Iran
[3] School of Engineering, Qom University, Qom, Iran
[4] Iran University of Science and Technology, Tehran, Iran
Emails: {hjuzi, rabieizadeh, ebaraty, bminaei}@noornet.net

Abstract

Nowadays similarity detection is one of the most applicable aspects of text mining techniques. There are different methods for similarity detection. This paper presents a new system for text similarity detection in Islamic Large Hadith Corpus of Computer Research Center of Islamic Science (CRCIS). This system uses N-gram method and Cosine measure for similarity detection. According to evaluation result, computer-based similarity detection systems can be more efficient than the previous related work in text similarity detection. We have obtained a 97% F-Score of similarity detection for Hadith texts. We hope that our system enables researches to find the unified Hadiths and to detect that how one large Hadith is divided into several small pieces of Hadith in different traditional Hadith books. This system would be very fruitful for many researches in the area of Hadith and the Holy Qur'an investigations.

Keywords: Text similarity, similarity detection, Hadith, prophetic traditions, text mining in Islamic texts

1. Introduction

Text similarity detection techniques followed many applications, including: plagiarism detection systems, content classifying and query-based information retrieval. Religious texts have been processed as one of the textual data groups by various text mining processes. Processing religious texts in Islam is more important than other religions since on the one hand the universality of Islam has caused the narration of a remarkable number of Hadiths[1] about all aspects of human life; on the other hand, Islam -as the last heavenly religion[2] was born nearly to the contemporary period of other religions. So, there are more and better techniques and tools for writing and keeping Hadiths.

The background of similarity detection in Islamic texts dates back to 1,000 years ago when the Muslim scholars began to check the authenticity of Hadiths and to classify them topically. To do so, they applied the Hadiths similarity detection manually and produced linkage between similar Hadiths.

Some of the most important functions of the Hadiths similarity detection are: to find the earliest reference of a Hadith; to find the genuine version of a Hadith; to compare the chains of Hadiths narrators with one another in order to remove available ambiguities; to retrieve the books missing during the history by collecting all of its citations in other references.

A number of activities have been done so far in the field of Hadiths Similarity detection. Fouzi Harrag et al. (Harrag & Hamdi-Cherif, 2007) for instance, worked on a system that classifies the Hadiths narrated from Prophet Mohammad (P. B. U. H) on the basis of the user's query. Khitam Jbara (Jbara, 2010) and Mohammed Naji Al-Kabi (Al-Kabi & Sinjilawi, 2007) worked on Hadith classification. Mohammed Q. Shatnawi (Shatnawi, Abuein, & Darwish, 2011) proposed a system that extracts Hadith texts from web pages and evaluates their authenticity.

2. Proposed System

2.1. Preprocessing Phase

Preprocessing is very important in the proposed system. Using linguistics and artificial intelligence techniques to prepare Hadith texts in this phase has

[1] That is the words, an accounts of the acts of Prophet Mohammad (P. B. U. H)

[2] Year 622 A.D

direct effect on the quality of the system results, the increase of the speed of the system and its precision (Ceska & Fox, 2009).

Each Hadith consist of two parts. The first part is the chain of Hadith narrators and the second part is the Hadith text. In this phase we removed the first part. All preprocessing tasks were applied on the Hadith text.

The following is the main preprocessing tasks of our system:

2.1.1. Normalization and Tokenization

The first step in preprocessing phase is normalization of the Hadith texts. In this step characters with more than one written form are changed to a unique form. For example, characters 'أ' and 'إ' are replaced with 'ا'. Also numbers and punctuation marks are removed to improve the results (Xu, Fraser, & Weischedel, 2002).

After normalization of spaces in the Hadith texts, the full space character are used for tokenization.

2.1.2. Removing Stop Words

Stop words are words with no effective meaning, (like some of particles, relative nouns, etc.). Removing stop words reduces the total words is a text, the words vector size, speed up the system and improves the quality and precision of the system results.

2.2. Analysis and Similarity Detection Phase

In this phase the prepared Hadith texts are divided into a sequence of n-grams (Laila, 2006). We examine n-grams with the length of one, two, and three characters. Experiment result shows that the best output is obtained when n-grams with length of three characters are used (Barron-Cedeno & Rosso, 2009).

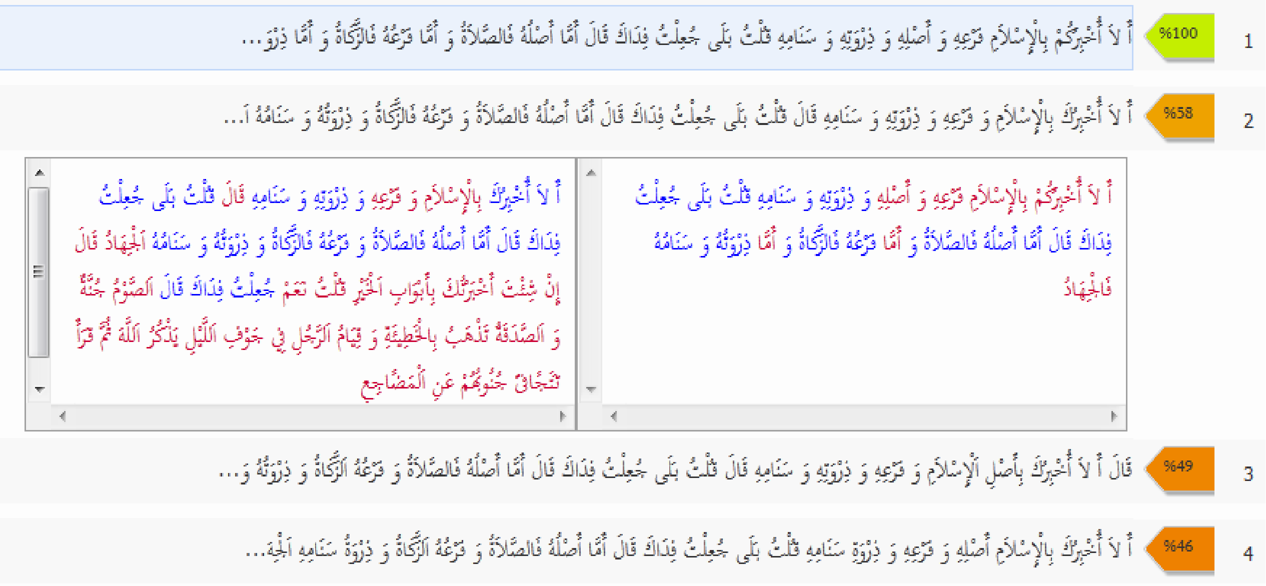
To increase precision of system output and to weight the n-gram sequences we use the following TF and IDF formulas (SALTON & BUCKLEY, 1988):

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

Then each Hadith is compared with all other Hadiths in the corpus and then the similarity is calculated. We tested different similarity measures like: Cosine, Dice and Jaccard and found that the Cosine measure provides better results (Al-Kabi & Sinjilawi, 2007). We followed the Cosine formula below:

$$cosine(P1, P2) = \frac{\sum_{i=1}^{n} P1_i \times P2_i}{\sqrt{\sum_{i=1}^{n} (P1_i)^2} \times \sqrt{\sum_{i=1}^{n} (P2_i)^2}}$$



1 %100 أ لاَ أُخْبِرُكُمْ بِالْإِسْلاَمِ فَرْعِهِ وَ أَصْلِهِ وَ ذِرْوَتِهِ وَ سَنَامِهِ قُلْتُ بَلَى جُعِلْتُ فِدَاكَ قَالَ أَمَّا أَصْلُهُ فَالصَّلاَةُ وَ أَمَّا فَرْعُهُ فَالزَّكَاةُ وَ أَمَّا ذِرْوَ...

2 %58 أ لاَ أُخْبِرُكَ بِالْإِسْلاَمِ وَ فَرْعِهِ وَ ذِرْوَتِهِ وَ سَنَامِهِ قَالَ قُلْتُ بَلَى جُعِلْتُ فِدَاكَ قَالَ أَمَّا أَصْلُهُ فَالصَّلاَةُ وَ فَرْعُهُ فَالزَّكَاةُ وَ ذِرْوَتُهُ وَ سَنَامُهُ ا...

أ لاَ أُخْبِرُكُمْ بِالْإِسْلاَمِ فَرْعِهِ وَ أَصْلِهِ وَ ذِرْوَتِهِ وَ سَنَامِهِ قُلْتُ بَلَى جُعِلْتُ
فِدَاكَ قَالَ أَمَّا أَصْلُهُ فَالصَّلاَةُ وَ أَمَّا فَرْعُهُ فَالزَّكَاةُ وَ أَمَّا ذِرْوَتُهُ وَ سَنَامُهُ
فَالْجِهَادُ

أ لاَ أُخْبِرُكَ بِالْإِسْلاَمِ وَ فَرْعِهِ وَ ذِرْوَتِهِ وَ سَنَامِهِ قَالَ قُلْتُ بَلَى جُعِلْتُ
فِدَاكَ قَالَ أَمَّا أَصْلُهُ فَالصَّلاَةُ وَ فَرْعُهُ فَالزَّكَاةُ وَ ذِرْوَتُهُ وَ سَنَامُهُ الْجِهَادُ قَالَ
إِنْ شِئْتَ أَخْبَرْتُكَ بِأَبْوَابِ الْخَيْرِ قُلْتُ نَعَمْ جُعِلْتُ فِدَاكَ قَالَ الصَّوْمُ جُنَّةٌ
وَ الصَّدَقَةُ تَذْهَبُ بِالْخَطِيئَةِ وَ قِيَامُ الرَّجُلِ فِي جَوْفِ اللَّيْلِ يَذْكُرُ اللَّهَ ثُمَّ قَرَأَ
تَتَجَافَى جُنُوبُهُمْ عَنِ الْمَضَاجِعِ

3 %49 قَالَ أ لاَ أُخْبِرُكَ بِأَصْلِ الْإِسْلاَمِ وَ فَرْعِهِ وَ ذِرْوَتِهِ وَ سَنَامِهِ قَالَ قُلْتُ بَلَى جُعِلْتُ فِدَاكَ قَالَ أَمَّا أَصْلُهُ فَالصَّلاَةُ وَ فَرْعُهُ الزَّكَاةُ وَ ذِرْوَتُهُ وَ...

4 %46 أ لاَ أُخْبِرُكَ بِالْإِسْلاَمِ أَصْلِهِ وَ فَرْعِهِ وَ ذِرْوَةِ سَنَامِهِ قُلْتُ بَلَى جُعِلْتُ فِدَاكَ قَالَ أَمَّا أَصْلُهُ فَالصَّلاَةُ وَ فَرْعُهُ الزَّكَاةُ وَ ذِرْوَةُ سَنَامِهِ الْجِهَ...

Figure 1: The system output example

Implemented system is accessible for everybody on the web[3]. As an example we show the system

[3] http://labs.noornet.net/SimilarHadith.aspx

output for following Hadith:

قال النبی (ص): "أَلاَ أُخْبِرُكُمْ بِالْإِسْلاَمِ فَرْعِهِ وَ أَصْلِهِ وَ ذِرْوَتِهِ وَ
سَنَامِهِ قُلْتُ بَلَى جُعِلْتُ فِدَاكَ قَالَ أَمَّا أَصْلُهُ فَالصَّلاَةُ وَ أَمَّا فَرْعُهُ
فَالزَّكَاةُ وَ أَمَّا ذِرْوَتُهُ وَ سَنَامُهُ فَالْجِهَادُ"

The system output for this Hadith shows 14 Hadiths with different similarity percentage (Figure 1). As shown in this figure, the words that do not exist in the similar Hadith found are specified in a different color.

2.3. Introducing the Applied Hadith Corpus

The Hadith Corpus used in this system contains in fact, an enormous set of rich Islamic Hadith books compiled since early Islamic centuries and have been digitalized by Computer Research Center of Islamic Science[4] (CRCIS) during 15 years. All this books have been gathered together in Jami`al-Ahadith application.[5] Some specifications of this corpus are showed in table 1.

Total Number of Books	308
Total Number of Volumes	630
Total Number of Hadiths	401,683
Number of Hadiths under 200 characters length	200,092
Total Number of Words	27,335,437
Number of Distinct Words	609,752
Total Number of Distinct Words After Preprocessing	234160
Total Number of Distinct trigrams	14005398

Table 1: The used corpus specification

3. Evaluation

To evaluate our system we used the Precision, Recall (Baeza-Yates & Ribeiro-Neto, 1999) and follow the F-Measure formulas.

$$Precision = \frac{number\ of\ relevant\ hadiths\ found}{number\ of\ hadiths\ found}$$

$$Recall = \frac{number\ of\ relevant\ hadiths\ found}{number\ of\ relevant\ hadiths\ to\ find}$$

$$F\_Measure = \frac{2PR}{P+R}$$

We selected 200 Hadiths randomly and inputted them to the system and gave them to a Hadith expert who has about 20 years of experience in Hadith evaluation. The precision of our system was 98%.

To calculate the recall of our system, we used Wasa'il al-Shi'ah which is a rich Hadith book containing more than 35,000 hadiths on Islamic rulings compiled by Sheikh Horr Ameli (d. 1104/1692) . The author has manually collected more than 35,000 hadiths from different references. Then he classified them in different groups each containing similar hadiths. This book is in 30 volumes the first of which contains 3,000 Hadiths. We selected the first 100 Hadiths of this volume. The selected Hadiths and their similar ones formed a set of 400 Hadiths. Then we inputted these 100 Hadiths to our system and compared the output with the similar Hadiths in that book. The recall of our system was 96%.

The final F-Measure was 96.98%. Our system proposed some similar Hadiths not mentioned in that book. In this case the Hadith expert is to evaluate all of outputs with similarity percentage of 70% and above.

4. Conclusion

In this paper, a new system for similarity detection in Islamic Hadith texts has been introduced. This system used some important preprocessing tasks, dividing texts into n-grams and using Cosine similarity measure. The precision of system was 98%, the recall was 96% and the F-Measure was 96.98%.

In future, we hope to test the following techniques to improve similarity detection process in Islamic Hadith texts:

Keyword extraction techniques, POS taggers, Removing words affices, lemmatization (Menai, 2011), name entity recognition (Zhang & Tsai, 2009), wordnet (Chien-Ying, Yeh, & Ke, 2010) and ontologies.

5. References

Al-Kabi, M. N., & Sinjilawi, S. I. (2007). A COMPARATIVE STUDY OF THE

[4] http://noorsoft.org/index.php?newlang=english

[5] http://noorsoft.org/modules.php?name=Contenten&pa=showpageen&pid=104

EFFICIENCY OF DIFFERENT MEASURES TO CLASSIFY ARABIC TEXT. *University of Sharjah Journal of Pure and Applied Sciences 4 (2)*, 13-26.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. *Addison Wesley/ACM Press*.

Barron-Cedeno, A., & Rosso, P. (2009). On Automatic Plagiarism Detection Based on n-Grams Comparison. *Advances in Information Retrieval -Springer*.

Ceska, Z., & Fox, C. (2009). The Influence of Text Pre-processing on Plagiarism Detection. *Proceedings of the International Conference RANLP*.

Chien-Ying, C., Yeh, J.-Y., & Ke, H.-R. (2010). Plagiarism Detection using ROUGE and WordNet. *Arxiv preprint*.

Harrag, F., & Hamdi-Cherif, A. (2007). UML modeling of text mining in Arabic language and application to the prophetic traditions “Hadiths”. *1st Int. Symp. on Computers and Arabic Language*.

Jbara, K. (2010). Knowledge Discovery in Al-Hadith Using Text Classification Algorithm. *Journal of American Science*.

Laila, K. (2006). Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. *Conference on Data Mining| DMIN*.

Menai, B. E. (2011). APlag: A plagiarism checker for Arabic texts. *Computer Science & Education (ICCSE), 6th International Conference on* , 1379 - 1383.

SALTON, G., & BUCKLEY, C. (1988). TERM-WEIGHTING APPROACHES IN AUTOMATIC TEXT RETRIEVAL. *Information processing & management*.

Shatnawi, M. Q., Abuein, Q. Q., & Darwish, O. (2011). Verification Hadith Correctness in Islamic Web Pages Using Information Retrieval Techniques. *ICICS*.

Xu, J., Fraser, A., & Weischedel, R. (2002). Empirical Studies in Strategies for Arabic Retrieval. *SIGIR Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* .

Zhang, Y., & Tsai, F. S. (2009). Combining named entities and tags for novel sentence detection. *ESAIR Proceedings of the WSDM Workshop on Exploiting Semantic Annotations in Information Retrieval* .

A Greek-Chinese Interlinear of the New Testament Gospels

John Lee[1], Simon S. M. Wong[2], Pui Ki Tang[1] and Jonathan Webster[1]
[1]The Halliday Centre for Intelligent Applications of Language Studies, City University of Hong Kong
[2]United Bible Societies
E-mail: {jsylee,pktang,ctjjw}@cityu.edu.hk, ssmwonghk@yahoo.com

Abstract

This paper describes an interlinear text consisting of the original Greek text of the four gospels in the New Testament, and its Chinese gloss. In addition to the gloss, the Greek text is linked to two linguistic resources. First, it has been word-aligned to the *Revised Chinese Union Version*, the most recent Chinese translation of the New Testament; second, it has been annotated with word dependencies, adapted from dependency trees in the PROIEL project. Through a browser-based interface, one can perform bilingual string-based search on this interlinear text, possibly combined with dependency constraints. We have evaluated this interlinear with respect to the accuracy, consistency, and precision of its gloss, as well as its effectiveness as pedagogical material.

1. Introduction

A bilingual corpus typically consists of a source text and its translation in a foreign language. In addition to these two parallel texts, an interlinear text (or, simply, "an interlinear") provides linguistic information for each word in the source text, most commonly a morphological analysis and a gloss, i.e., a translation of the word in the foreign language. The typical user of an interlinear is primarily interested in the source text, and needs linguistic support to read it.

Interlinears can serve readers at every level of proficiency of the source language. Advanced students can enjoy a quicker and smoother pace of reading, since the glosses reduce the interruptions caused by dictionary look-ups. For beginners, even dictionaries can be difficult to use, since the conversion from the inflected form of a word to its root form, or dictionary form (e.g., in Greek, from the surface form *andra* to the root form *anēr*), is no trivial task; choosing the appropriate meaning from the various senses provided in the dictionary entry (e.g., *anēr* can mean 'man' or 'husband') can also be challenging. Morphological analyses and glosses remove both obstacles.

For readers who do not aim to learn the source language, interlinears can still reveal important linguistic features of the source text that are obscured in the foreign translation. For example, Koine Greek explicitly places the word *heis* 'one' in front of a noun to emphasize the uniqueness of the entity in question (Bauer et al., 2001); unfortunately, the equivalent Chinese word, *yi* 'one', cannot convey this emphasis since it is routinely used with all nouns to indicate indefiniteness. A Chinese reader, if made aware of the presence of the word *heis*, would better appreciate the idea of 'one and only' that is strongly underlined by the Greek author.

This paper describes a Greek-Chinese interlinear of the four gospels --- Matthew, Mark, Luke and John --- which constitute the first four books of the New Testament in the Bible. The rest of the paper is organized as follows. Section 2 reviews previous work. Section 3 outlines the design principles of our interlinear. Section 4 discusses the implementation of the interlinear. Section 5 presents evaluations in section 5, followed by conclusions.

2. Previous Work

While interlinears have a long history in opening up Classical works to modern readers (Ovid, 1828; Xenophon, 1896), they are also valued for religious texts, whose readers attach special reverence for sacred texts in their original languages. For example, both the Qur'an (Eckmann, 1976; Dukes and Habash, 2010; etc.) and the Bible have been interlinearized with various languages (Green, 1984; Wang, 1985; etc.)

For the New Testament, most interlinear texts have been in English (Green, 1984; Marshall, 1993; Brown and Comfort, 1993); so far, only one Greek-Chinese version has been published (Wang, 1985). This pioneer work is, however, almost completely outdated from the point of view of both the Chinese and Greek texts.

For the Chinese text, Wang (1985) used the *Chinese Union Version* [1] (1919). Since the glosses were also selected mainly from this 90-year-old translation, their vocabulary, expressions and word senses diverge considerably from contemporary Chinese usage. For the Greek text, Wang (1985) used the 1952 edition of the *Novum Testamentum Graece* (Nestle and Aland, 1952), a version that is no longer acceptable to the scholarly community.

Our interlinear brings both texts up-to-date with the latest Greek edition (Nestle and Aland, 1994) and the recent *Revised Chinese Union Version* (2010). Henceforth, we refer to these as the "Greek text" and the "Chinese text".

[1] This translation in turn reflects a 19th-century edition of the Greek New Testament (Palmer, 1881).

3. Interlinear Design

Our interlinear improves upon (Wang, 1985) in two ways. In terms of *new content*, we indicate correspondences to Greek morphology in the Chinese glosses (section 3.1), and provide word alignments between the Greek text and the Chinese text (section 3.2); in terms of *methodology*, we address issues in gloss accuracy, precision and consistency (sections 3.3 to 3.5).

3.1 Correspondences to Greek morphology in Chinese glosses

To indicate case, gender, number, tense or mood, Greek employs an extensive system of suffixes that are attached to the lemmas of nouns, adjectives and verbs. Chinese expresses the equivalent information not with suffixes, but with separate characters. It is critical to convey these concepts to the Chinese speaker; to this end, our Chinese glosses clearly distinguish between characters corresponding to the meaning of the Greek lemma, and those corresponding to the suffix(es).

Greek	***boēthei***	***kyrie***
	‘help!’	‘O Lord’
Chinese gloss in our interlinear	求你#幫助	主#啊
	qiu ni#bangzhu	*zhu#a*
	‘beg you#help’	‘Lord#O’
Chinese gloss in (Wang, 1985)	幫助	主啊
	bangzhu	*zhu a*
	‘help’	‘O Lord’

Table 1. Two example glosses illustrating how our Chinese glosses reflect Greek morphology. The pound sign in each gloss separates the meaning of the lemma from that of the suffix. On the left, the characters *qiu ni* ‘beg you’ before the pound sign expresses the imperative mood of the verb *boēthei*. In the example on the right, the character *a* expresses the vocative case of the noun *kyrie*. In contrast, the glosses in (Wang, 1985) do not distinguish these meanings.

For instance, as shown in Table 1, the Chinese gloss for the Greek word *boēthei* ‘help!’, in the imperative mood, contains a pound sign that separates the characters *bangzhu* ‘help’ (meaning of the Greek lemma) from *qiu ni* ‘beg you’ (meaning of the imperative mood). Similarly, the gloss for *kyrie* ‘O Lord’ has two clearly demarcated components, *zhu* ‘Lord’, and *a* ‘O’, the latter of which expresses the vocative case. In contrast, the glosses in (Wang, 1985) mix the meaning of the suffix with that of the lemma (e.g., *zhu a* for *kyrie*), and sometimes even omit the former (e.g., simply *bangzhu* for *boēthei*).

3.2 Word Alignment

Our interlinear is digitally searchable, allowing readers to retrieve Greek word(s) equivalent to a Chinese word, and vice versa. In the two verses in Figure 1, the word *patēr* ‘father’ in the Greek text is shown to correspond to the word 父 *fu* in the Chinese text, and both *ēgapēsen* ‘loved’ and *philei* ‘loves’ correspond to 愛 *ai* ‘love’.

This search function critically depends on word alignments between the two texts. It is inadequate to rely on the Chinese gloss: the search would fail whenever the gloss deviates from the exact wording in the Chinese text. We therefore specify direct word alignments between the Greek and Chinese texts.

Another benefit of these alignments is to connect the Chinese text to the gloss. The connection is sometimes not evident in non-literal translations; for example, at first glance, the word 迎娶 *yingqu* ‘marry’ in the Chinese text does not obviously correspond to 在一起 *zaiyiqi* ‘be together’, the gloss for *synelthein* ‘to be together’. Given the alignment between *yingqu* and *synelthein*, the reader can see that the word for ‘marry’ in the Chinese text in fact translates a word that literally means ‘to be together’ in the original Greek text.

3.3 Gloss Accuracy

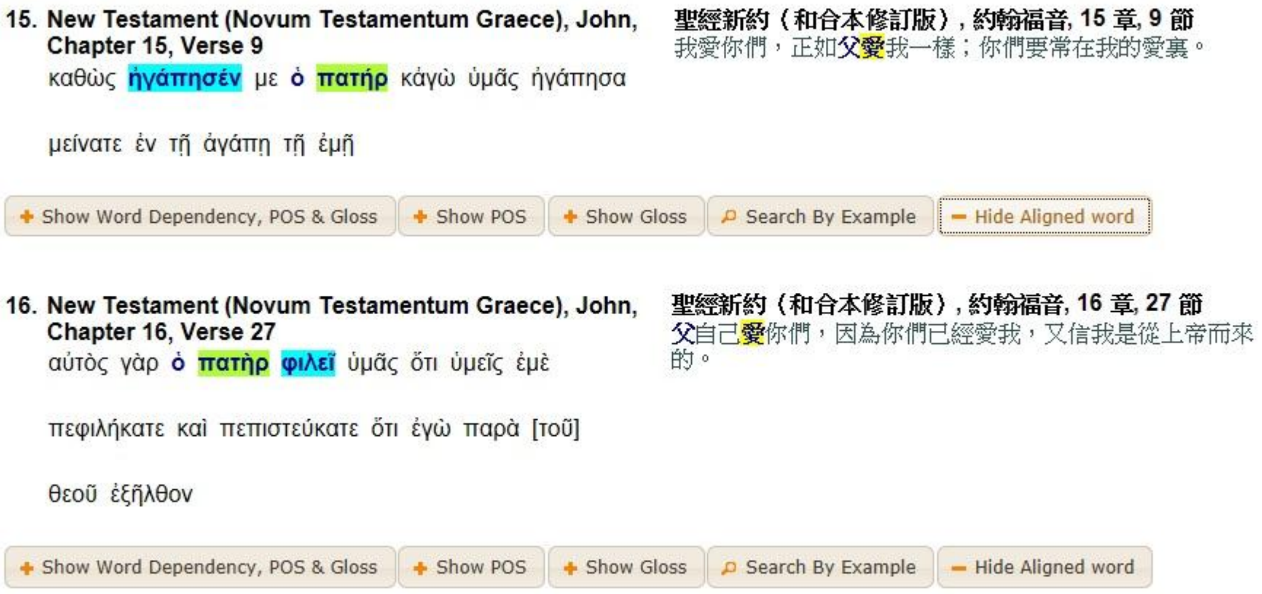
In addition to the two types of new content described above, our interlinear also makes advances in gloss accuracy, precision and consistency. An accurate gloss is one that expresses the exact meaning of the word in question, without mixing with other words in the sentence.

The treatment of expressions involving conjunctions and prepositions leaves much room for improvement in (Wang, 1985). First, the glosses for verbs often contain superfluous conjunctions, such as 就 *jiu*, 而 *er*, and 便 *bian* ‘then’. For instance, *anexōrēsen* ‘departed’ was glossed as 就退 *jiu tui* ‘then departed’ (Matthew 2:22), *egeneto* ‘became’ as 而成 *er xing* ‘then became’ (John 1:3, 1:10, 1:17), and *eipen* ‘said’ as 便告訴 *bian gaosu* ‘then said’ (Matthew 3:7). These conjunctions suggest meanings that are not at all present in the Greek, and they risk confusing the reader. Although they might improve the reading fluency, especially when the glosses were to be read as a sentence, the sacrifice in accuracy is hardly justifiable.

A second problematic construction is the Greek prepositional phrase (PP), which is often expressed by two Chinese words that are non-contiguous. Consider the PP *epi gēs* ‘on earth’, which means 在地上 *zai di shang* ‘at earth top-of’. The preposition *epi* ‘on’ corresponds not only to *zai* ‘at’, but rather to *zai ... shang* ‘at … top-of’. The tendency in (Wang, 1985), unfortunately, is to follow the Chinese word order, often at the expense of accuracy. The PP above is a typical case: *epi* was inaccurately glossed as 在 *zai* ‘at’; while *gēs* ‘earth’, which means only *di* ‘earth’, was assigned 地上 *di shang* ‘earth top-of’, the remainder of the equivalent Chinese PP.

3.4 Gloss Precision

Besides being accurate, the gloss should also be precise. Precision means distinguishing the finer shades of meaning of a word, including its implied meaning, where appropriate.



15. New Testament (Novum Testamentum Graece), John, Chapter 15, Verse 9

καθὼς ἠγάπησέν με ὁ πατήρ κἀγὼ ὑμᾶς ἠγάπησα

μείνατε ἐν τῇ ἀγάπῃ τῇ ἐμῇ

聖經新約（和合本修訂版）, 約翰福音, 15 章, 9 節

我愛你們，正如父愛我一樣；你們要常在我的愛裏。

16. New Testament (Novum Testamentum Graece), John, Chapter 16, Verse 27

αὐτὸς γὰρ ὁ πατὴρ φιλεῖ ὑμᾶς ὅτι ὑμεῖς ἐμὲ

πεφιλήκατε καὶ πεπιστεύκατε ὅτι ἐγὼ παρὰ [τοῦ]

θεοῦ ἐξῆλθον

聖經新約（和合本修訂版）, 約翰福音, 16 章, 27 節

父自己愛你們，因為你們已經愛我，又信我是從上帝而來的。

Figure 1. Two of the verses retrieved from a search of the Chinese word 愛 *ai* 'love' aligned to a Greek verb with a noun subject, based on dependency annotations from PROIEL. In the top verse, the Greek verb is *ēgapēsen* 'loved', in the bottom one, it is *philei* 'loves'; the subject for both is *patēr* 'father'. These results demonstrate that the Greek language distinguishes between two kinds of love that are not reflected in the Chinese, and also list the entities that can be the agents of these kinds of love.

The most frequently occurring word in the Greek text, *kai*, generally means 'and' or 'then', but also has a wide semantic range including 'but', 'also', 'even', and sometimes bears no particular meaning (Wallace, 1997). In (Wang, 1985), *kai* was indiscriminately given the gloss 就 *jiu* 'then' or 而[且] *er[qie]* 'and'. Our glosses, in contrast, operate at a finer level of granularity, further distinguishing instances where the context calls for 但 *dan* 'but', 卻 *que* 'but', 又 *you* 'also', 甚至 *shenzhi* 'even', or <empty>.

It is helpful for the gloss to indicate not only the literal meaning of a word, but also its implied meaning. Sometimes it is a matter of restoring grammatical features missing in Greek, such as indefinite articles. To illustrate, while *huios* 'son' can be rendered simply as 兒子 *erzi* 'son', it is more precise to include the Chinese number and counter word, i.e., (一個)兒子 *(yige)erzi* '(one) son'. In other cases, the implied meaning can be a narrowing or an expansion of the literal meaning. For instance, *adelphē* generally means 'sister', for which 姐妹 *jiemei* 'sister' would suffice; but in some contexts, its meaning is clearly limited to 'Christian sister', and it should instead be glossed as (信主的)姐妹 *(xinzhude) jiemei* '(believing) sister'. In the opposite direction, *adelphos* generally means 'brother', but when the context calls for the expanded meaning 'brothers and sisters', it is more precise to use the gloss 兄弟(姐妹) *xiongdi(jiemei)* 'brothers (and sisters)'.

3.5 Gloss Consistency

A gloss can be both accurate and precise, but not consistent; i.e., when two synonymous but different glosses are used for two words in the source text that have the exact same meaning. The treatment in (Wang, 1985) of the phrase *episteusan eis auton* 'believed in him' is a case in point. Among its seven occurrences in the Gospel of John, *episteusan* 'believed' was variously glossed as 信了 *xinle*, 相信 *xiangxin*, and 就相信了 *jiuxiangxinle*, all near-synonyms meaning 'believed'; *eis* 'in' was glossed in three ways, namely, 深入 *shenru* 'enter deeply into', 歸順 *guishun* 'submit to', and 歸入 *guiru* 'return to'. The mix-and-match of these different glosses can easily mislead the reader to the conclusion that the Greek phrase had many different nuanced meanings. To facilitate consistency, we use Paratext, a tool developed at United Bible Societies. This software keeps track of all Chinese glosses that have been used for each Greek word, and lists them in descending frequency. The translator consults this list, and re-uses an existing translation where possible. In our interlinear, *episteusan* is consistently glossed as 相信 *xiangxin* 'believed', except when it bears another sense, such as in Luke 16:11, where an alternate gloss 託付 *tuofu* 'entrust' is in order.

4. Implementation

The source text of our interlinear consists of about 65,000 Greek words. Their glosses were produced with help from a specialized software (section 4.1). The interlinear has been enhanced with Greek word dependencies, and can be searched on a browser (section 4.2).

4.1 Interlinearization Software

Paratext is a software developed by the United Bible Societies to help translators, exegetical advisors and consultants to produce quality translations from the point of view of both format and content. The translator can first input a first or revised draft of the text, then check

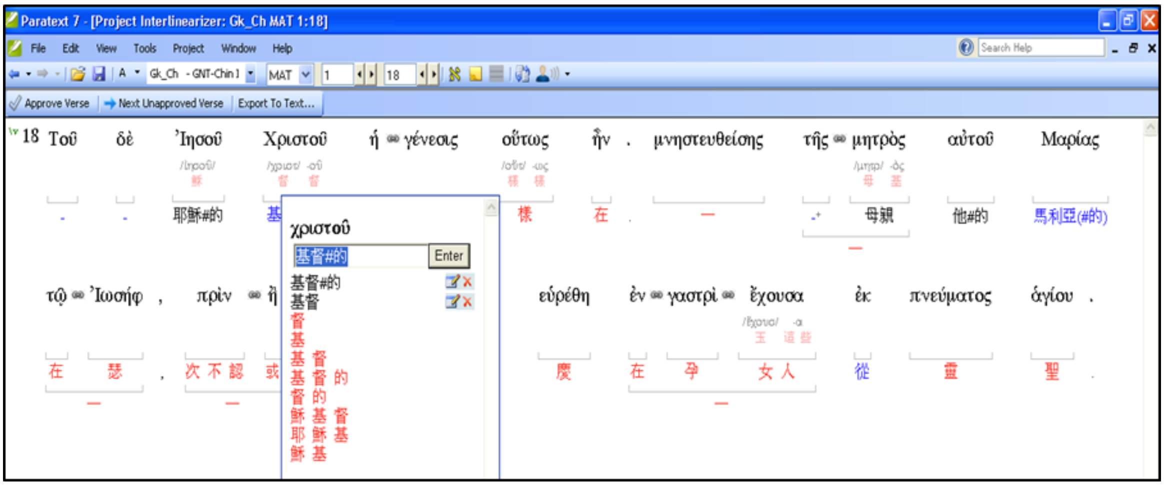
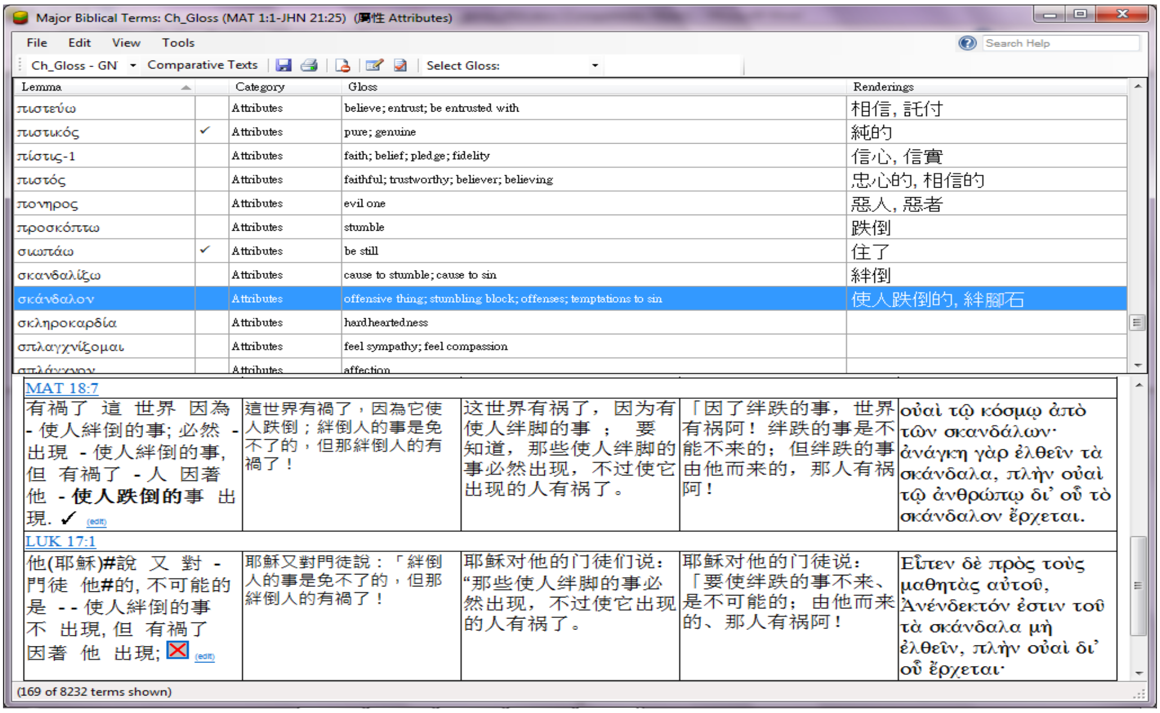

Paratext 7 - [Project Interlinearizer: Gk_Ch MAT 1:18]
File Edit View Tools Project Window Help
Search Help
Gk_Ch - GNT-Chin1
MAT 1 18
Approve Verse
Next Unapproved Verse
Export To Text...
18 Τοῦ δὲ Ἰησοῦ Χριστοῦ ἡ γένεσις οὕτως ἦν . μνηστευθείσης τῆς μητρὸς αὐτοῦ Μαρίας
/Ἰησοῦ/ 穌
/χριστ/ -οῦ 督 督
/οὕτ/ -ως 樣 樣
/μητρ/ -ὸς 母 基
- - 耶穌#的 基 樣 在 . 一 -+ 母親 他#的 馬利亞(#的)
一
χριστοῦ
基督#的
Enter
基督#的
基督
督
基
基 督
基 督 的
督 的
穌 基 督
耶 穌 基
穌 基
τῷ Ἰωσήφ , πρὶν ἢ εὑρέθη ἐν γαστρὶ ἔχουσα ἐκ πνεύματος ἁγίου .
/ἔχουσ/ -α 玉 這 些
在 瑟 , 次 不 認 或 慶 在 孕 女 人 從 靈 聖 .
一 一 一

Figure 2. Interface of the “Project Interlinearizer” tool in Paratext.



Lemma		Category	Gloss	Renderings
πιστεύω		Attributes	believe; entrust; be entrusted with	相信, 託付
πιστικός	✓	Attributes	pure; genuine	純的
πίστις-1		Attributes	faith; belief; pledge; fidelity	信心, 信實
πιστός		Attributes	faithful; trustworthy; believer; believing	忠心的, 相信的
πονηρος		Attributes	evil one	惡人, 惡者
προσκόπτω		Attributes	stumble	跌倒
σιωπάω	✓	Attributes	be still	住了
σκανδαλίζω		Attributes	cause to stumble; cause to sin	絆倒
σκάνδαλον		Attributes	offensive thing; stumbling block; offenses; temptations to sin	使人跌倒的, 絆腳石
σκληροκαρδία		Attributes	hardheartedness	
σπλαγχνίζομαι		Attributes	feel sympathy; feel compassion	
σπλάγχνον		Attributes	affection	

MAT 18:7				
有禍了 這 世界 因為 - 使人絆倒的事; 必然 - 出現 - 使人絆倒的事, 但 有禍了 - 人 因著 他 - **使人跌倒的事** 出 現. ✓ (edit)	這世界有禍了，因為它使人跌倒；絆倒人的事是免不了的，但那絆倒人的有禍了！	这世界有祸了，因为有使人绊脚的事 ； 要知道，那些使人绊脚的事必然出现，不过使它出现的人有祸了。	「因了绊跌的事，世界有祸阿！ 绊跌的事是不能不来的；但绊跌的事由他而来的，那人有祸阿！	οὐαὶ τῷ κόσμῳ ἀπὸ τῶν σκανδάλων· ἀνάγκη γὰρ ἐλθεῖν τὰ σκάνδαλα, πλὴν οὐαὶ τῷ ἀνθρώπῳ δι’ οὗ τὸ σκάνδαλον ἔρχεται.
LUK 17:1				
他(耶穌)#說 又 對 - 門徒 他#的, 不可能的 是 - - 使人絆倒的事 不 出現, 但 有禍了 因著 他 出現; ☒ (edit)	耶穌又對門徒說：「絆倒人的事是免不了的，但那絆倒人的有禍了！	耶稣对他的门徒们说："那些使人绊脚的事必然出现，不过使它出现的人有祸了。	耶稣对他的门徒说：「要使绊跌的事不来、是不可能的；由他而来的、那人有祸阿！	Εἶπεν δὲ πρὸς τοὺς μαθητὰς αὐτοῦ, Ἀνένδεκτόν ἐστιν τοῦ τὰ σκάνδαλα μὴ ἐλθεῖν, πλὴν οὐαὶ δι’ οὗ ἔρχεται·

(169 of 8232 terms shown)

Figure 3. Interface of the “Biblical Terms” tool in Paratext. The top panel shows the list of important Greek biblical terms, their English gloss and Chinese renderings; the bottom panel shows the glosses assigned to verses where the terms appear.

that draft against the biblical source texts, and against a selection of model translations and resource materials in electronic format.

We made use of two important features in Paratext. Shown in Figure 2, “Project Interlinearizer” is a tool that can create an auto-generated (back) translation of any text in an interlinear view, based on statistical analyses of the glosses. Among the many applications of this tool in biblical translation work, the drafting of a Greek-Chinese interlinear draft is certainly a natural one. In the gloss review process (section 5.1), an important functionality of this tool is to record the gloss(es) selected for each word.

4.2 Search interface

Our interlinear can be accessed via a browser-based search interface, which offers various retrieval options: with Chinese words, Greek words and lemmas, or Greek syntactic information, or any combination thereof.

Word dependencies. Dependency trees of the Greek New Testament from the PROIEL project (Haug & Jøhndal, 2008) have been adapted and integrated into this interface. This treebank not only provides grammatical dependencies between the Greek words in the gospels, but also their lemma and morphological parsing.

The treebank is based on the 8th edition of the Greek New Testament by Tischendorf (1869 --- 1872), which has a considerable number of differences with our Greek text (Nestle and Aland, 1994). We aligned the two versions, and amended the trees where there are discrepancies.

Search interface. On the browser-based interface, the user can specify up to five terms, and then further constrain their relations. A term can be either a Chinese word, or a Greek word, lemma, or a part-of-speech tag from the PROIEL tagset. A relation can be syntactic, i.e., a head-child relation between two Greek terms. For example, to see what entities can serve as the subject of the verb *philei* 'loves', one can search for all occurrences of *philei* that is connected to a child noun with the dependency label "sub". Alternatively, a relation can be a word alignment between a Chinese word and a Greek term. Figure 1 shows results from a search that uses both languages as well as dependency information from PROIEL.

5. Evaluation

Our interlinear has been evaluated with respect to its gloss quality (section 5.1) and pedagogical effectiveness (section 5.2).

5.1 Gloss Evaluation

After an initial draft, we reviewed the Chinese glosses using a feature called "Biblical Terms" in Paratext. Shown in Figure 3, this tool assists translators and consultants in reviewing key terms of the translated text, based on the source text. About 1000 words long, the list of key terms include attributes, beings, animals, plants, objects, rituals, names and miscellaneous. For each term, we determine its principal senses, consulting if necessary the authoritative dictionary *A Greek-English Lexicon of the New Testament and Other Early Christian Literature* (Bauer et al., 2001). To check collocation units, especially recurring and idiomatic expressions, phrases from (Aland, 1983) have also been selected to complement this list.

We performed a detailed analysis on gloss precision, accuracy and consistency on a small subset of these terms. Bauer et al. (2001) organize the meaning of a word in a hierarchy, and provide, for each meaning, example verses from the New Testament illustrating its use. For each term, we constructed its 'synsets' (i.e., its word senses) according to the top-level categories in this dictionary, and retrieved the Chinese gloss assigned to the Greek word in each of the example verses cited. We then measured the following:

- ***Consistency***: In principle, each sense should be rendered with one gloss. This metric asks to what extent Chinese glosses vary within a synset (but does not judge the appropriateness of the glosses);
- ***Accuracy***: A gloss should reflect the meaning of the corresponding Greek word. This metric asks whether the Chinese glosses in a synset express an appropriate meaning (even if their wordings differ);
- ***Precision***: A gloss should be fine-grained enough to distinguish between major word senses, therefore the same gloss should normally not be used for words in two different synsets. This metric measures how often this occurs.

Nine Greek words were chosen for this analysis based on two criteria: their frequency in the gospels, and amount of word-sense ambiguity. Each of these words straddles at least two, and up to four, synsets. Altogether, 329 instances of their use are cited as examples by Bauer et al. (2001), and are examined in this analysis[2].

The average gloss accuracy is 98%. Almost all of the inaccurate cases are due to interpretations. For instance, *sēmeion* generally means 'sign', 'token', or 'indication', but in one instance it is glossed as 對象 *duixiang* 'target'. This is an interpretation beyond the word's surface meaning given in the dictionary, but is supported by at least one mainstream Chinese translation of the New Testament.

The average consistency is 77%. Often, the presence of multiple Chinese glosses in one synset was caused by a mismatch of the level of granularity with Bauer et al. (2001). To illustrate, three different glosses 預兆 *yuzhao* 'sign', 記號 *jihao* 'token' and 神蹟 *shenji* 'miracle' were found in one synset for *sēmeion*. These nuances all belong to the same top-level category in the dictionary entry of *sēmeion*, but are in fact further differentiated as subcategories.

Word	**Accuracy**	**Consistency**	**Precision**
archē	100%	63%	100%
alētheia	100%	92%	87%
hamartōlos	100%	94%	62%
epitimaō	100%	66%	100%
krinō	100%	45%	81%
logos	98%	95%	96%
peithō	100%	86%	100%
sēmeion	95%	91%	91%
psychē	94%	62%	70%
Average	*98%*	*77%*	*87%*

Table 2. Evaluation results on the Chinese gloss accuracy, consistency and precision for nine common Greek words.

The average precision is 87%. For some terms, subtle differences in shades of meaning were difficult to distinguish, resulting in the application of the same Chinese gloss in multiple synsets. This phenomenon is most frequent for the adjective *hamartōlos*, which, according to Bauer et al. (2001), can be used substantively to mean either 'a sinner' or 'an irreligious person'. The

[2] The total number of their occurrences in the Gospels is, of course, greater than 329.

difference is often not straightforward to discern in the New Testament context, and the Chinese gloss uses the Chinese word 罪人 *zuiren* 'sinner' for both. Similarly, the noun *psychē* can mean 'soul' or 'life'. As acknowledged by Bauer et al. (2001), some instances of *psychē* in the 'soul' synset may also serve as a metaphor for 'life'. In a few of these cases we did use the gloss 生命 *shengming* 'life', further contributing to a lower precision.

5.2 Pedagogical Evaluation

We have deployed this interlinear in the course "Elementary Ancient Greek" at our university, with the goal to expose students to authentic Greek texts as early as possible. Twelve students, with no previous knowledge of Greek, were enrolled in the course. At the first two lectures, they were introduced to the Greek alphabet and pronunciation. The third lecture presents the concepts of case and definiteness, for which there are no Chinese equivalents. At the following lecture, students learned the forms of the definite article and adjectives of the first and second declensions; our interlinear was used at this point to teach the following two topics.

Adjectival constructions. In Greek, an attributive adjective can appear after an article but in front of the noun[3] ("adjective-first"); alternatively, it can also follow the noun, in which case both the noun and the adjective have an article[4] ("noun-first"). Rather than directly teaching these constructions, the instructor asked students to induce them from our interlinear, by examining the occurrences of a handful of paradigm adjectives that they had just learned.

With help from the Chinese glosses, the students independently identified the noun modified by the adjective in each instance, and observed the relative positions of the article, adjective and noun. Eleven out of the 12 students were able to formulate the two possible constructions, citing examples such as *ton kalon oinon* ('the good wine', in "adjective-first" construction) and *ho poimēn ho kalos* ('the good shepherd', in "noun-first" construction).

Verb endings. In a second exercise, the instructor asked students to induce the verb endings for the present indicative active forms, by examining a large number of occurrences of the verb *legō* 'say', and separating the stem from the ending of the verb. As in the previous exercise, eleven out of the 12 students successfully hypothesized the endings, and specified the person (first, second or third) and number (singular or plural) to which the each ending corresponds.

In both exercises, the Chinese glosses and word alignments played a critical role in the learning experience. Since their vocabulary was very limited, the students relied on the glosses to locate the Greek adjectives and verbs, as well as the relevant linguistic context, e.g., articles in the first exercise and noun subjects in the second. The Chinese information was thus indispensable in enabling these beginners a direct encounter with authentic Greek text barely a month after their introduction to the language.

6. Conclusion

We have reported the development and evaluation of a Greek-Chinese interlinear text of the gospels in the Greek New Testament. Based on the most current texts available for both languages, it emphasizes gloss accuracy, precision and consistency, and contains new linguistic information such as word alignments and correspondences to Greek morphology in the Chinese gloss. A Greek dependency treebank has been adapted for the text. A search interface, offering bilingual string-based retrieval with dependency constraints, has been developed and successfully deployed in a Greek language class.

7. References

Aland, K. (1983). *Vollstaendige Konkordanz zum Griechischen Neuen Testament unter zugrundelegun aller modernen kritischen textausgaben und des Textus Receptus*. Berlin: Walter de Gruyter.

Bauer, Walter, and Danker, Frederick William (2001). *A Greek-English Lexicon of the New Testament and Other Early Christian Literature*. Chicago: University of Chicago Press.

Brown, Robert K. and Comfort, Philip W. (1993). *The New Greek-English Interlinear New Testament*. Tyndale House Publishers.

Chinese Union Version 和合本 (1919). Hong Kong: Hong Kong Bible Society.

Dukes, Kais and Habash, Nizar (2010). Morphological Annotation of Quranic Arabic. In *Proc. LREC*.

Eckmann, J. (1976). *Middle Turkic Glosses of the Rylands Interlinear Koran Translation*. Budapest: Publishing House of the Hungarian Academy of Sciences.

Green, Jay P. (1984). *The Interlinear Greek-English New Testament, with Strong's Concordance Numbers Above Each Word*. Peabody: Hendrickson Publishers.

Haug, Dag and Jøhndal, Marius (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proc. LREC Workshop on Language Technology for Cultural Heritage Data*.

Marshall, Alfred (1993). *The Interlinear NASB-NIV Parallel New Testament in Greek and English*. Grand Rapids: Zondervan.

Nestle, Erwin and Kurt Aland et al. (1952). *Novum Testamentum Graece*. 21st edition. Stuttgart: Deutsche Bibelgesellschaft

Nestle, Erwin and Kurt Aland et al. (1994). *Novum Testamentum Graece*. 27th edition. Stuttgart: Deutsche Bibelgesellschaft.

Ovid (1828). *The first book of Ovid's Metamorphoses,*

[3] i.e., in the sequence article-adjective-noun
[4] i.e., in the sequence article-noun-article-adjective

with a literal interlinear translation and illustrative notes. London: University of London.

Palmer, Edwin (1881). *ΚΑΙΝΗ ΔΙΑΘΗΚΗ - The Greek Testament with the Readings Adopted by the Revisers of the Authorised Version*. Oxford: Clarendon Press.

Revised Chinese Union Version 和合本修訂版 (2010). Hong Kong: Hong Kong Bible Society.

Wallace, Daniel B. (1997). *Greek Grammar Beyond the Basics*. Grand Rapids: Zondervan.

Wang, C. C. (1985). *Chinese-Greek-English Interlinear New Testament*. Taichung: Conservative Baptist Press.

Xenophon (1896). *The first four books of Xenophon's Anabasis: the original text reduced to the natural English order, with a literal interlinear translation*. Harrisburg: Handy Book Corp.

Linguistic and Semantic Annotation in Religious *Memento mori* Literature

Karlheinz Mörth, Claudia Resch, Thierry Declerck, Ulrike Czeitschner

Austrian Academy of Sciences – Institute for Corpus Linguistics and Text Technology (ICLTT)

Sonnenfelsgasse 19/8, A-1010 Vienna

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Stuhlsatzenhausweg, 3, D-66123 Saarbrücken

E-mail: karlheinz.moerth@oeaw.ac.at, claudia.resch@oeaw.ac.at, declerck@dfki.de, ulrike.czeitschner@oeaw.ac.at

Abstract

The project described in this paper was at first concerned with the specific issue of annotating historical texts belonging to the *Memento mori* genre. To produce a digital version of these texts that could be used to answer the specific questions of the researchers involved, a multi-layered approach was adopted: Semantic annotations were applied to the digital text corpus in order to create a domain-specific taxonomy and thus to facilitate the development of innovative approaches in both literary and linguistic research. In addition, the project aimed to develop text technological methodologies tailored to this particular type of text. This work can be characterised by a high degree of interdisciplinarity, as research has been carried out with both a literary/historical and linguistic/lexicographic perspective. The annotations created as part of this project were also designed to be used in the adaptation of existing linguistic computational tools to suit the needs of non-canonical language varieties.

Keywords: annotation, standards, historical corpora

1. Introduction

The research presented in this paper is based on a growing digital collection of printed German language texts dating from the Baroque era, in particular the years from 1650 until 1750. The specific type of text represented in the corpus consists of religious writings largely motivated by the fear of sudden death: *Memento mori* (or "Be mindful of dying") ideas proliferated during this time period and confrontation with Death was often evoked or at least encouraged by the church. Clergymen spread their beliefs by preaching, counselling and offering advice. Their religious books were meant to remind the audience of the fragility of their existence and the futility of human ambition. By reading these moralising texts, people were supposed to be directed towards holding themselves in steady expectation of their own demise and admonished to live a life of virtue.

Books of religious instruction concerning "last things" had a central place in Baroque culture, saw a number of reprinted editions and were a best-selling genre during the 17th century. However, this type of text has more or less been neglected by scientific research (Dreier, 2010; Wunderlich, 2000), which can be partially explained by the fact that many of these writings are housed in monastic libraries and not easily available to interested researchers.

2. A Corpus of Austrian Early Modern German

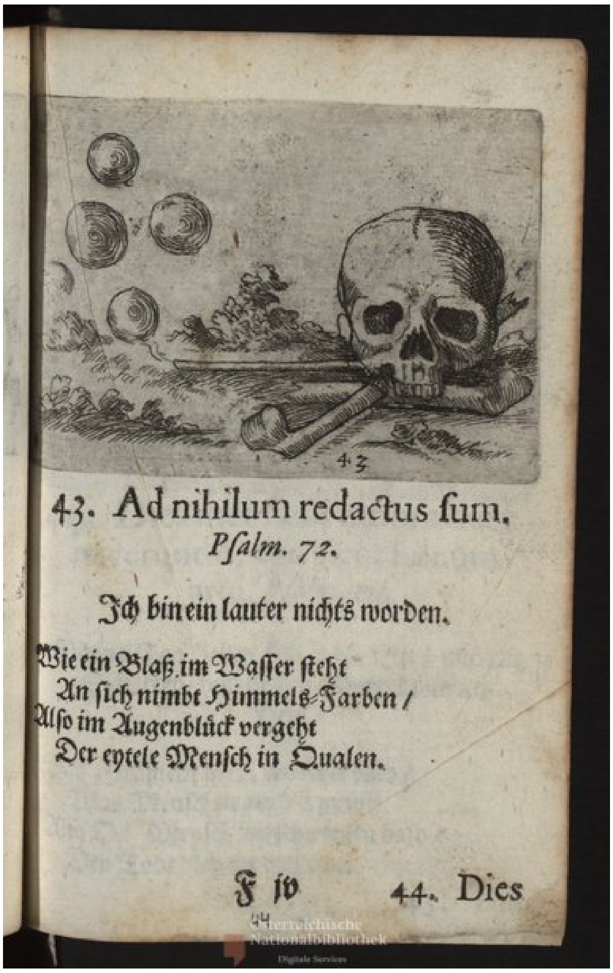
Our project has produced a digital corpus of complete texts and images belonging to the *Memento mori* genre. The collection presently consists of a comparatively small number of complete versions (not just samples) of first editions of such devotional books in prose and verse yielding some 150.000 running words. Many of the books included in the corpus are richly illustrated with copperplate-prints. While the majority of the selected works can be clearly ascribed to the Baroque Catholic writer Abraham a Sancta Clara (1644-1709)[1], the question of the authorship of parts of the corpus could not be satisfactorily resolved. Because of his literary talent and peculiar style, the discalced preacher was very popular in Vienna, which was zealously pious under the reign of the Habsburg emperors at that time. "The fact that Abrahams books sold well, led several publishers to the idea of combining parts of his already published works with the texts of other authors, ascribing the literary hybrid to the Augustinian preacher." (Šajda, 2009).[2] Other texts of the corpus can be positively attributed to authors who were part of the clergyman's social environment such as his uncle[3], other friars of the religious order of the pious brotherhood[4], his publishers and imitators.

[1] The publications of Franz M. Eybl (2008 and 2011) contain biographical information about Ulrich Megerle, who later became known as Abraham a Sancta. In 1662 he joined the order of Discalced Augustinians, which had formed a community in Vienna in 1631. In the following years he held several leading positions in his order.

[2] This could be the case with his allegedly "last" text entitled "Besonders meublirt- und gezierte Todten-Capelle / Oder Allgemeiner Todten-Spiegel" (1710) and several others.

[3] Abraham Megerle: Speculum musico-mortuale. Salzburg 1672.

[4] E.g. Klare / vnd Warhaffte Entwerffung / Menschlicher Gestalt / vnd Wesenheit […] Gedruckt zu Wienn […] im Jahr 1662. The "Bruderschaften" of the Baroque era were Church authorized societies which had the religious perfection of particular devotions, worship practices, and acts of charity as their goal. Abraham a Sancta Clara was the spiritual father of such a society and published several fraternal books which are also part of the digital collection.



43. Ad nihilum redactus ſum.

Pſalm. 72.

Jch bin ein lauter nichts worden.

Wie ein Blaß im Waſſer ſteht
An ſich nimbt Himmels=Farben /
Alſo im Augenblück vergeht
Der eytele Menſch in Qualen.

F jv 44. Dies

Figure 1: Image of one of the pages collected in the digital corpus.

```
<pb facs="n0087.jpg"/>
<div type="chapter">
<figure/>
<head type="main" rend="antiqua"><seg
type="enum">43.</seg>
<cit><quote><foreign xml:lang="la">Ad
nihilum redactus
sum.</foreign></quote><lb/>
<bibl rend="italicised">Psalm.
72.</bibl></cit></head>
<head type="sub">Jch bin ein lauter
nichts worden.</head>
<lg><l>Wie ein Blaß im Wasser steht</l>
<l>An sich nimbt Himmels=Farben /</l>
<l>Also im Augenblück vergeht</l>
<l>Der eytele Mensch in Qualen.</l></lg>
</div>
<fw place="bot_center" type="pageNum">F
jv</fw>
<fw place="bot_right" type="catch">44.
Dies</fw>
```

Figure 2: Example of a textual raw XML-data.

3. Applied Technologies

The corpus is encoded using TEI (P5)[5]. The mark-up applied has been designed to capture several layers of information: (a) fundamental structural units such as paragraphs, headings, line groups, etc. (much attention has been paid to annotating structural, typographical and orthographical details), (b) word class information and lemmas, (c) named entities (place names, personal names including biblical figures), (d) biblical quotations, and (e) salient semantic features of terms personifying death.

Part of the project has been dedicated to methods of dealing with this particular brand of language, which differs considerably from the modern standard varieties of German with respect to lexis, morphology and syntax. In addition, all writings from that time display a remarkable degree of orthographic variation, which makes the application of natural language processing approaches difficult. In a first step, standard tools were used to add basic linguistic information (POS, lemmas) to the texts. In a second step, this data was manually proofread and corrected. One of the goals of this project is the creation of a digital dictionary of the linguistic variety under investigation, which in turn will be used to improve the performance of existing natural language processing tools.

4. Semantic Annotation of ‘Death’-Related Concepts

The most salient topic in these religious texts is ‘death and dying’, which is dealt with in a most inventive manner by modifying the image of the so-called ‘reaper figure’, one of the most apparent personifications of death.

To study the ways in which these notions were encoded into texts and in order to allow adequate decoding of their semantics, all ‘death-related’ lexical units (including metaphors and figures representing ‘Death’) were first semi-automatically marked-up by means of a shallow taxonomy, which was further developed over the course of the annotation process. We are currently working on methods to optimise and eventually automatise this process. One tool, which we are testing to improve the results of the annotation process, is NooJ, a freely available linguistic development environment.[6] This tool allows us to combine the lexicographic resources extracted from the corpus and grammatical rules to parse new additions to the corpus to achieve improved results.

A preliminary statistical evaluation proved that, of all the nouns, *Tod* (or ‘death’ in English), including deviant spellings such as „Todt”, was most commonly used. In terms of frequency, it is closely followed by words such as *Gott* ‘God’, *Herr* ‘Lord’, *Heil* ‘salvation’ and *Menschen* ‘human beings’. One particular feature of the genre is that ‘death’ is often characterised as a dark, bony character

[5] See http://www.tei-c.org/Guidelines/P5/ for more details.

[6] http://www.nooj4nlp.net/pages/nooj.html
The format for lexical encoding in NooJ is supporting the description of variants of lemma. A reason why we opted for this platform.

carrying a large scythe, a sickle or bow and arrows. The personification of death as an omnipresent ruler was an invention of the later Middle Ages, but writers and preachers of the Baroque period, like Abraham a Sancta Clara, were extremely creative in modifying this image of the *reaper figure*, which was the most apparent personification of death and has been preserved in art, literature and popular culture to the present day.

5. Encoding details

The encoding was designed to capture not only nominal simplicia such as *Tod, Todt* or *Mors*, but also compound nouns (*Aschen=Mann* ‘man made of ash’, *Rippen-Kramer* ‘huckster who deals in ribs’) and other nominal multi-word expressions (*General Haut und Bein* ‘General Skin and Bones’). Another category includes attributes of ‘death’, such as *knochenreich* ‘rich in bones’, *tobend* ‘raging’ or *zaun-dürr* ‘spindly thin’ and verbal phrases such as *nehm ich auch beim Schopff* ‘I also seize by a tuft of hair (i.e. I kill)’. The following list of terms furnishes ample evidence for the richness of the death-related vocabulary of this early variety of New High German.

Aschen=Mann | Auffsetzer | Blumen=Feind | Bock | Bücher=Feind | Capell=Meister | Caprioln=Schneider | Contre Admiral Tod | Bereuter | Dieb | Herr Doctor Tod | Donnerkeil | Dürrer | Fieber=Tinctur | Fischer | Gall=Aepffel=Brey | Gärtner | Gefreitter | General Haut und Bein | Gesell | Grammaticus | Haffner | Hechel= und Maus=Fallen Jubilirer | Herr von Beinhausen und Sichelberg | Holtzhacker | Kercker=Meister aller gefangenen Seelen | Knoll | Kräuter=Salat | Lebens=Feind | Mader | Mann | Marode-Reuter | Menschen=Fischer | Menschen=Mörder | Menschen=Schnitter | Menschen=Würger | Mißgeburth | Oppositum der Schnecken=Post | Schützen=Meister | Pedell | Perforce-Jäger | Reuter auf dem fahlen Pferd | Rippen=Kramer | Schnitter | Schütz | Sensentrager | Soldaten=Feind | Spielmann | Spieler | Strassenrauber | Tantz=Meister | Tertius | Töpffer | Waghalse | Weltstürmer

Figure 3: Death-related nouns documented in the corpus

The research on this topic aims to establish a detailed taxonomy. We plan to offer a semantic representation of the particular vocabulary, which can then be re-used in diachronic as well as synchronic linguistic studies of similar lexical fields.

6. Outlook

By making use of the guidelines of the Text Encoding Initiative for this small historical corpus with a specialized research focus (scil. the representation of death in religious texts), we aim to achieve a high level of interoperability and re-usability of the developed annotations so that the technologies successfully applied to these Baroque era texts might also be useful for the analysis of other historical as well as contemporary texts written in non-canonical language varieties.

It is important to mention that our research interests are not only focused on literary studies and semantic annotations, the data is also being used by the ICLTT’s tech crew for the development of a prototypical web-interface. This tool, called corpus_shell, has been designed as a generic web-based publication framework for heterogeneous and physically distributed language resources. The components of this system will be made freely accessible as part of the ICLTT’s engagement in the CLARIN-AT and DARIAH-AT infrastructure projects.

The Corpus of Austrian Early Modern German will partially go online in 2012. The digital texts will become available together with the facsimiles of the original prints and related metadata; and users will have open access to the underlying sources.

As digital language resources from the Baroque era are scarce (the Viennese corpus is one of the very few collections[7] containing data from this period), each additional text constitutes a valuable contribution to the field of study. With this in mind, the corpus is currently expanding and has also acquired obituaries, religious chants and funeral sermons from this period. In addition to the quantitative expansion of the corpus, a cooperation with a research group[8] currently annotating other text genres from the same historical period has been established, and combining resources and results is benefitting both projects mutually.

7. References

Bennett, P., Durrell, M., Scheible, S., Whitt, R. J. (2010): *Annotating a historical corpus of German: A case study.* Proceedings of the LREC 2010 workshop on Language Resources and Language Technology Standards. Valletta: pp. 64--68.

Boot, P. (2009): *Mesotext. Digitised Emblems, Modelled Annotations and Humanities Scholarship*. Amsterdam: University Press.

Czeitschner, U., Resch, C. (2011): *Texttechnologische Erschließung barocker Totentänze*. In U. Wunderlich (Ed.) *L’art macabre. Jahrbuch der Europäischen Totentanz-Vereinigung.* Bamberg.

Declerck, T., Czeitschner, U., Mörth, K., Resch, C., Budin, G. (2011). *A Text Technology Infrastructure for Annotating Corpora in the eHumanities*. In S. Gradmann, F. Borri, C. Meghini & H. Schuldt (Eds.), *Research and Advanced Technology for Digital Libraries*. Berlin: Springer, pp. 457--460.

Dipper, S. (2010): *POS-Tagging of Historical Language Data: First Experiments.* In *Semantic Approaches in Natural Language Processing.* Proceedings of the 10th

[7] E. g. projects like the *GerManC project* (University of Manchester) http://www.llc.manchester.ac.uk/research/projects/germanc/ or parts of the *Bonner Frühneuhochdeutschkorpus* http://www.korpora.org/Fnhd/.

[8] Grimmelshausen Corpus (University of Heidelberg in cooperation with Herzog August Bibliothek Wolfenbüttel).

Conference on Natural Language Processing (KONVENS-10). Saarbrücken: pp. 117--121.

Dreier, R. P. (2010). *Der Totentanz – ein Motiv der kirchlichen Kunst als Projektionsfläche für profane Botschaften (1425-1650).* Leiden: Printpartners Ipskamp.

Eybl, F. M. (2008). *Abraham a Sancta Clara*. In Killy Literaturlexikon Volume 1. Berlin: de Gruyter, pp. 10--14.

Eybl, F. M. (to appear). *Abraham a Sancta Clara. Vom barocken Kanzelstar zum populären Schriftsteller.* In A. P. Knittel (Ed.) *Beiträge des Kreenheinstetter Symposions anlässlich seines 300. Todestages*. Eggingen.

Greule A., Kucharska-Dreiss, Elzbieta (Eds.) (2011). *Theolinguistik. Bestandsaufnahme – Tendenzen – Impulse*. Volume 4. Insingen: Bauer & Raspe.

Reffle, U. (2011). *Efficiently generating correction suggestions for garbled tokens of historical language*. In *Natural Language Engineering*, 17 (2). Cambridge University Press, pp. 265--282.

Šajda, P. (2009): *Abraham a Sancta Clara: An Aphoristic Encyclopedia of Christian Wisdom*. In *Kierkegaard and the Renaissance and Modern Traditions – Theology*. Ashgate.

Wunderlich, U. (2000). *Zwischen Kontinuität und Innovation – Totentänze in illustrierten Büchern der Neuzeit*. In *„Ihr müßt alle nach meiner Pfeife tanzen“, Totentänze vom 15. bis 20. Jahrhundert aus den Beständen der Herzog August Bibliothek Wolfenbüttel und der Bibliothek Otto Schäfer Schweinfurt.* Wolfenbüttel.

The Qur’an Corpus for Juz’ Amma

Aida Mustapha[1], Zulkifli Mohd. Yusoff[2], Raja Jamilah Raja Yusof[2]
[1]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia
[2]Centre of Quranic Research, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
E-mail: aida@fsktm.upm.edu.my, zulkifliy@um.edu.my, rjry@um.edu.my

Abstract

This paper presents a corpus that offers rich knowledge for Juz’ Amma. The corpus is designed to be in dual-language, which are English and Malay. The knowledge covers translation for each word and verse, tafsir, as well as hadith from different authenticated sources. This corpus is designed to support dialogue interaction with an information visualization system for Quranic text called AQILAH. This corpus is hoped to aid mental visualization in studying the Qur’an and to enable the users to communicate the content of Juz’ Amma with clarity, precision, and efficiency.

Keywords: Quranic corpus, Juz’ Amma, Dialogue system

1. Introduction

Contrary to book presentation that is systematic and coherent, with each subject is definite and neatly divided into sections and chapters, the Qur’an presents the subject of Truth, belief and conduct, good tidings or condemnation in alternate manner without any apparent system. The same subject is dealt over and over again across different chapters throughout the Holy Book. In light with this very specific nature of the Qur’an, studying Qur’an requires a comprehensive, if not redundant, source that would nourish more than a superficial understanding of the Holy Book.

Juz’ Amma is the thirtieth chapter in the Qur’an. This last chapter consists of 37 surah from *An-Naba* (Surah 78) to *An-Naas* (Surah 114). There are 564 verses (ayat) altogether from all the surahs with varying length. 34 surahs in this juz’ were revealed in Makkah and the remaining four were revealed in Madinah. Most of the surahs also contain the earliest revelations. These early surahs are primarily concerned with oneness of Allah, day of judgement and the afterlife.

The corpus for Juz’ Amma has been developed with two objectives. First is to contribute to a rich knowledge base sourced from different authenticated Muslim scholars. Second is to serve dialogue-based information visualization system for Quranic text called AQILAH (Mustapha, 2009). Juz’ Amma is chosen as the pilot study because it contains the greatest number of surahs of any juz’ and the surahs are most commonly recited and memorized.

AQILAH is an ambitious effort to build a Quranic visualization system with the objective to qualitatively enhance human experience in navigating Quranic text using human-machine dialogues. This is in line with the finding that use of natural language dialogues via conversational agents offer higher understanding of the presented information (Beun et al., 2003). AQILAH facilitates human-computer interaction and rearrange the Quranic content based on need-to-know basis, as opposed to sequential order as in recitals. The system is equipped with a minimal technology of simple keyword-based parsing to parse text-based input of human natural language. It responds in natural language based on the knowledge provided by this corpus, by means of extracting keywords from an input string and performing keyword matching against the knowledge base.

The following Section 2 reviews some literature related to Quranic text, Section 3 reports the corpus design, and Section 4 concludes with some direction for future work.

2. Related Work

Researches on Arabic language at both syntactic and semantic level are growing. However, only a small percentage of such research is sourced from the Quranic text. Among researches on Quranic text at syntactic level include work related to building a dependency Treebank (Dukes and Buckwalter, 2010; Dukes et. al, 2010), word stemming (Yusof et al., 2010), morphological annotation (Dukes and Habash, 2010), and prosody (Younis, 2011).

Meanwhile, at the semantic level, Sharaf (2009) perform deep annotation and analysis of the Qur’an by means of semantic modeling of the knowledge contained in each verse. In ontology development, Saad et al. (2009; 2010; 2011) focus on ontology creation for Islamic concepts such as solat as sourced from the Quranic text. Others include application of rules extraction (Zaidi et al., 2010) and semantic lexicons in ontology development (Al-Yahya et al., 2010).

At pragmatic and discourse level, Sharaf and Atwell (2009) is the first known work to design a knowledge representation model for the Qur’an based on the concept of semantic net. Subsequent works include knowledge acquisition from Quranic text (Saad et al., 2011) and text categorization (Sharaf and Atwell, 2011). While all existing work study the entire Qur’an, the aim for this work is modest, which is to prepare a working Quranic knowledge base scoped to only Juz’ Amma.

3. Corpus Design

The objective of this corpus design is to prepare rich knowledge source for Juz’ Amma on the basis of individual verse or ayat. To achieve this, each ayat in this corpus is annotated with literal translation, tafsir, and supporting hadith. The most important feature in this corpus is the keyword for each ayat that is manually extracted based on semantic content of the ayat.

3.1 Translation

The main source of the English translation for the Qur’an is based on the book The Noble by Hilali and Khan (2007). While there are a number of translations widely available, Hilali-Khan’s translation is chosen as the main source of translation because it provides a more comprehensive interpretation of the Qur’an based on the original understanding the Prophet’s sayings and teachings. The explanation of the meaning and interpretations of The Noble Qur’an is also based on tafsir of Ibn Jarir At-Tabari, Al-Qurtubi, and Ibn Kathir as well as from the hadith from Prophet Muhammad s.a.w. as mentioned in the authentic hadith reported by Imam Bukhari.

The second source is from modern English translation by Abdel Haleem (2005) in his book The Qur’an. Unlike other translations including the Hilali-Khan’s who use old English or literal translation for Arabic words, this translation is written in contemporary language, yet remains faithful to the meaning and spirit of the original text. It avoids literal translation especially idioms and certain local phrases in Arabic. In terms of depth of translation, translation by Abdel Haleem provides geographical and historical notes, and keeps the commentary and explanation in the footnotes. This makes the translation more readable and easier to understand instead of using parenthesis as in Hilali-Khan's.

Translation in the Malay language is sourced from Tafsir Pimpinan Al-Rahman (Basmeih, 2001) published by the Department of Islamic Development Malaysia (JAKIM). Because it is endorsed by the local authority, this translation has been widely used throughout Malaysia and is often used for preaching Islam. The translation refers to authenticated tafsir and hadith resources such as by Al-Tabari, Ibn Kathir, Sayyid Qutub and many other reknown Muslim scholars. This translation is easy to understand and capitalizes on footnotes to elaborate on certain Arabic concepts.

3.2 Tafsir

Tafsir (or tafseer) is a body of commentary that aims to explain the meaning of verses in the Qur’an. The English tafsir in this corpus is sourced from the tafsir by Sayyid Abul Ala Mawdudi based on his book Towards Understanding the Qur’an (Mawdudi, 2007). This book offers a logical and effectively reasoned explanation for the tafsir and has been translated to English by Dr. Zafar Ishaq Ansari. Meanwhile, the tafsir in Malay comes from the book Tafsir Al-Qur’an Al-Hakim (Yusoff and Yunus, 2012). Note that tafsir books are never ayat-based, but more to lengthy explanation that spans over few sentences in explaining a particular ayat. Incorporating knowledge from tafsir is a laborious work as we have to read the entire surah before tagging it to the most suitable ayat.

3.3 Hadith

Hadith are sayings or oral traditions relating to words and acts by Prophet Muhammad s.a.w. Hadith collected for this corpus are also in dual language, which are English and Malay. However, the source of Hadith provided are not comprehensive but rather a collection of authenticated hadith as cited in both tafsir by Sayyid Mawdudi (2007) and Yusoff and Yunus (2012).

2.4 Keywords

The most important part of this corpus is annotation of one or more keywords to each verse (ayat) in Juz’ Amma. This work is performed manually because Qur’an vocabulary is very flexible and highly illustrative, depending on the scholar who translated it. The keyword extraction is performed by a Qur’an expert who read each verse in view of the entire context of particular translation, and assigns one or more keywords based on the translation.

Table 1 shows the difference in keywords as extracted from Surah At-Takwiir (The Overthrowing) verses 8-9 based on both translation by Hilali-Khan and Abdel Haleem.

Table 1: Keywords for verses 81:8 and 81:9.

Ayat	Hilali-Khan	Abdel Haleem
وَإِذَا ٱلۡمَوۡءُۥدَةُ سُئِلَتۡ ٨	the female (infant) buried alive (as the pagan Arabs used to do)	the baby girl buried alive
بِأَيِّ ذَنۢبٖ قُتِلَتۡ ٩	sin	sin

As shown in Table 1, keywords extracted from Hilali-Khan are more complicated as compared against to Abdel Haleem due to their style of writing in English translation.

The translation, tafsir, hadith, and keywords for each verse are prepared in a text file tagged with identifiers to indicate its purpose, such as hilali-khan, haleem, mawdudi, jakim, tafsir or hadith. Figure 1 and Figure 2 shows the extracted knowledge for ayat 8-9 Surah At-Takwiir (Surah 81) respectively.

Figure 1: Knowledge for verses 81:8.

ayat_id 081-08

keyword baby girl buried alive

hilali-khan And when the female (infant) buried alive (as the pagan Arabs used to do) is questioned

haleem And when the female infant buried alive is asked

mawdudi The case of the infant girl who was buried alive, should be decided and settled justly at some tune, and there should necessarily be a time when the cruel people who committed this heinous crime, should be called to account for it, for there was none in the world to hear the cries of complaint raised by the poor soul

jakim Dan apabila bayi-bayi perempuan yang ditanam hidup-hidup

tafsir Kejadian kelapan, anak-anak perempuan yang ditanam hidup-hidup sebagaimana yang dilakukan oleh masyarakat jahiliyyah disebabkan takut malu dan takut miskin

Based on Figure 1 and Figure 2, the tag `ayat_id` refers to specific ayat, for example `081-08` refers to Surah 81 (At-Takwiir) verse 8. For the English knowledge base, the tag `hilali-khan` refers to English translation by Hilali-Khan, `haleem` refers to English translation by Abdel Haleem, `mawdudi` is the English tafsir by Sayyid Mawdudi (2007). Finally, `hadith` is the English hadith sourced from his tafsir.

As for the Malay counterpart, `jakim` is the tag for Malay translation published by JAKIM, `tafsir` refers to the book by Yusoff and Yunus (2012), and `hadis` is a collection of hadis written in Malay as collected from the tafsir book by the same book of Yusoff and Yunus (2012).

Figure 2: Knowledge for verses 81:8.

ayat_id 081-09

keyword sin

hilali-khan For what sin, was she killed?

haleem For what sin she was killed

mawdudi The parents who buried their daughters alive, would be so contemptible in the-sight of Allah that they would not be asked: Why did you kill the innocent infant? But disregarding them the innocent girl will be asked: For what crime were you slain?

mawdudi And she will tell her story how cruelly she had been treated by her barbarous parents and buried alive)

hadith The Holy Prophet said to Suraqah bin Jusham: Should I tell you what is the greatest charity (or said: one of the greatest charities)? He said: Kindly do tell, O Messenger of Allah. The Holy Prophet said: Your daughter who (after being divorced or widowed) returns to you and should have no other bread-winner (Ibn Majah, Bukhari Al-Adab al-Mufrad)

hadith The Muslim who has two daughters and he looks after them well, they will lead him to Paradise (Bukhari: Al-Adab al-Mufrad)

hadith The one who has three daughters born to him, and he is patient over them, and clothes them well according to his means, they will become a means of rescue for him from Hell (Bukhari, Al-Adab al-Mufrad, Ibn Majah)

hadith The one who has a daughter born to him and he does not bury her alive, nor keeps her in disgrace, nor prefers his son to her, Allah will admit him to Paradise (Abu Daud)

jakim Kerana dosa apakah ia dibunuh

tafsir Bayi-bayi ini akan ditanya apa dosa sehingga ia ditanam dan siapa pembunuhnya dan hal ini merupakan satu ancaman kepada pembunuhnya

tafsir Hakikat pertanyaan kepada bayi maksudnya pertanyaan kepada orang yang melakukan perbuatan itu dan merupakan kejian serta kutukan kepada mereka

hadis Imam Ahmad telah meriwayatkan daripada Khansa binti Muawiyah al-Sarimiyyah daripada bapa saudaranya, katanya: Aku bertanya kepada Rasulullah s.a.w.: Wahai Rasulullah, siapakah yang masuk dalam syurga? Jawab Baginda: Nabi dalam syurga, para syuhada dalam syurga, anak-anak kecil yang mati sewaktu kecil dalam syurga dan anak yang hidup ketika masih kecil masuk syurga

4. Conclusion

The corpus developed for Juz' Amma is based on individual verse or ayat. However, ayat-based representation in this corpus is insufficient for dialogue navigation by AQILAH because the keywords are referenced to literal Arabic meaning, hence AQILAH will not be able to extract the most meaningful ayat for a given dialogue-based query. Furthermore, for every ayat, the number of semantic content associated with it i.e. translation, tafsir etc is not fixed. In light for this, ontology is a natural choice for knowledge representation for AQILAH. In the future, this corpus will be redesigned in the form of ontology to serve AQILAH.

In addition to AQILAH, the in-depth knowledge representation in this corpus can be used as a resource for other applications with regards to Quranic studies, for example in Question Answering System or Intelligent Tutoring System. It also provides a one-stop authentic and validated source of knowledge in both English and Malay language. It is hoped that this corpus is able to aid mental visualization in studying the Qur'an and to enable users to communicate the content of Juz' Amma with clarity, precision, and efficiency.

5. Acknowledgements

This research is funded by Fundamental Research Grants Scheme by Ministry of Higher Education Malaysia in collaboration with the Centre of Quranic Research at University of Malaya, Malaysia.

6. References

Abdel Haleem, M.A.S. (2005). *The Qur'an: A new translation*. New York: Oxford University Press.

Al-Yahya, M., Al-Khalifa, H., Bahanshal, A., Al-Odah, I. and Al-Helwa, N. (2010). An ontological model for representing semantic lexicons: An application on time nouns in the Holy Qur'an. Arabian Journal for Science and Engineering (AJSE).

Basmeih, S.A.M. (2001). Tafsir pimpinan Al-Rahman kepada pengertian Al-Qur'an. JAKIM, ISBN 9830420132.

Beun, R.J., Vos, E.d. and Witteman, C. (2003). Embodied conversational agents: Effects on memory performance and anthropomorphisation. In *Proceedings of the International Conference on Intelligent Virtual Agents*, Springer-Verlag, pp. 315--319.

Dukes, K. and Buckwalter, T. (2010). A dependency treebank of the Qur'an using traditional Arabic grammar. In *Proceedings of the 7th International Conference on Informatics and Systems* (INFOS). Cairo, Egypt.

Dukes, K., Atwell, E. and Sharaf, A.M. (2010). Syntactic annotation guidelines for the Quranic Arabic Treebank. In *Proceedings of Language Resources and Evaluation Conference* (LREC). Valletta, Malta.

Dukes, K. and Habash, N. (2010). Morphological annotation of Quranic arabic. In *Proceeding of the Seventh International Conference on Language Resources and Evaluation*.

Hilali, M.T. and Khan, M.M. (2007). *Translation of the meaning of the Noble Qur'an in the English language*. King Fahd Printing Compex, Madinah, Saudi Arabia, 956 pages. ISBN: 978-9960-770-15-4.

Mawdudi, S.A.A. (2007). *Tafhim Al-Qur'an : Towards understanding the Qur'an*. Islamic Foundation, ISBN 978-0860375104.

Mustapha, A. (2009). Dialogue-based visualization for Quranic text. *European Journal of Scientific Research* 37(1), pp. 36--40.

Saad, S., Salim, N. and Zainal, H. (2009). Islamic knowledge ontology creation. In *Proceedings of the International Conference for Internet Technology and Secured Transactions*, London, 9-12 November 2009, pp. 1--6.

Saad, S., Salim, N., Zainal, H. and Noah, S.A.M. (2010). A framework for Islamic knowledge via ontology representation. In *Proceedings of the 2010 International Conference on Information Retrieval and Knowledge Management*, Selangor, Malaysia, 17-18 March 2010, pp. 310--314.

Saad, S., Salim, N., Zainal, H. and Muda, Z. (2011). A process for building domain ontology: An experience in developing solat ontology. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, Bandung, Indonesia, 17-19 July 2011, pp. 1--5.

Saad, S., Salim, N. and Zainuddin, S. (2011). An early stage of knowledge acquisition based on Quranic text. In *Proceedings of the International Conference on Semantic Technology and Information Retrieval* (STAIR). Putrajaya, Malaysia.

Sharaf, A.M. (2009). *Knowledge representation in the Qur'an*. Interim Report, University of Leeds.

Sharaf, A. and Atwell, E. (2009) A corpus-based computational model for knowledge representation of the Qur'an. In *Proceedings of the Fifth Corpus Linguistics Conference*, Liverpool.

Sharaf, A. and Atwell, E. (2011). Automatic categorization of the Quranic chapters. In *Proceedings of the 7th International Computing Conference in Arabic* (ICCA11). Riyadh, Saudia Arabia.

Younis, N. (2011). *Semantic prosody as a tool for translating prepositions in the Holy Qur'an: A corpus-based analysis.* Workshop on Arabic Corpus Linguistics. Lancaster University.

Yusoff, Z. and Yunus, M. (2012) *Tafsir Al-Qur'an Al-Hakim*. Jilid 6, PTS Publication Sdn. Bhd.

Yusof, R., Zainuddin, R., Baba, M.S. and Yusof, Z. (2010). Qur'anic Words Stemming. *Arabian Journal for Science and Engineering* (AJSE).

Zaidi, S., Laskri. M. and Abdelali, A. (2010). Arabic collocations extraction using Gate. In *Proceedings of IEEE ICMWI'10*. Algiers, Algeria.

Lexical Correspondences Between the Masoretic Text and the Septuagint

Nathan Ellis Rasmussen and Deryle Lonsdale

Department of Linguistics & English Language
Brigham Young University
Provo, UT, USA 84602
volodymyr.velyky@gmail.com, lonz@byu.edu

Abstract

This paper describes a statistical approach to Biblical vocabulary, in which the parallel structure of the Greek Septuagint and the Hebrew Masoretic Text is used to locate correspondences between lexical lemmata of the two Biblical languages and score them with a log-likelihood ratio. We discuss metrics used to propose and select possible correspondences, and include an examination of twenty pre-selected items for recall and of twenty items just above the cutoff for precision. We explore the implications for textual correlation and translation equivalence.

Keywords: bitext, concordance, lexicon induction, log-likelihood, Masoretic Text, Septuagint

1. Background

The original text of the Bible is divided between Hebrew/Aramaic (hereafter, simply ‘Hebrew’) and Greek. Concordance studies in each of these languages have become a staple of Biblical interpretation over the past century and a half (Cotterell and Turner, 1989). However, the division between the two languages confines concordance studies to either the Old or the New Testament, limiting the testaments’ ability to interpret each other. Concordancing can expose collocation, semantic prosody, allusion, quotation, and other interpretive cues, but in order to use these across the intra-Biblical language barrier we must be sure our translated word is ‘loaded’ in more or less the same way as the original.

An objective link between the Biblical languages is offered by the Septuagint (hereafter LXX), an Old Testament translated from Hebrew to Greek about 300 B.C.E. and an important influence on New Testament language. Previous scholarship has noted that ‘there is hardly a verse in the [New Testament] the phraseology of which may not be illustrated, and to some extent explained, by reference to the LXX’ (Girdlestone, 1983). By comparing it statistically to the Hebrew Masoretic Text (MT) we can quantify the evidence in LXX for a given translation—‘extrapolate the mental dictionary of the translators’ (Joosten, 2002), a dictionary not only of univocal translation pairs, but of ‘subtleties . . . revealed primarily through examination of a variety of translated forms’ (Resnik et al., 1999).

Over a hundred key Hebrew words are matched to Greek equivalents in Robert Baker Girdlestone’s nineteenth-century *Synonyms of the Old Testament* (Girdlestone, 1983), purportedly via LXX, but he neither explains his methods nor provides citations. Therefore we must take his LXX data, his knowledge of the languages, and his traditions and opinions all together or not at all. ‘The Religious Vocabulary of Hellenistic Judaism’, first part of C. H. Dodd’s (1935) *The Bible and the Greeks*, gives thorough LXX citations, but covers a limited vocabulary—largely a subset of Girdlestone.

The nineteenth-century Hatch-Redpath concordance of the Septuagint (Hatch and Redpath, 1998) provides raw data in bulk, albeit from a text that predates accepted modern critical editions. It annotates each citation with its corresponding Hebrew word from MT, which makes Greek-to-Hebrew correspondence statistics available simply through counting. In the other direction, Hebrew to Greek, Muraoka (1988) indexes Hebrew words in Hatch-Redpath and provides all Greek equivalents—though without usage counts, which must be found by looking up each Greek word separately.

Muraoka’s ideal index ‘would have required study of every single verse of the Septuagint, comparing it with the extant Hebrew and Aramaic original texts’. This paper moves toward that ideal via the recent CATSS Parallel text (Tov, 2005).

This is a complete machine-readable parallel text of LXX and MT. Its alignments are primarily of one Hebrew word, with its clitics (prepositions, articles, some pronominal objects, and some conjunctions), to two or three words of Greek. It is freely available for academic use. This enables a replicable and thorough survey of Hebrew-Greek word pairs in MT and LXX—claiming support for each word pair in the handiwork of the ancient translators themselves.

Such is our approach to the content words of the Old Testament. We neglect personal and place names, since they may be expected to map one-to-one between languages. We omit functional words, which are poor objects of concordance studies because of their frequency and which interfered with our analysis in pilot studies.[1] We claim the advantage over Muraoka in accounting for ‘every single verse’ as he suggested, though he would also like to see data on textual variants and exhaustive hand analysis, both of which we defer.

Nor is the lemma-centric approach the limit of possibility. It misses the connotations of evocative pairings of words, the semantic and syntactic context crucial to a good model of prepositional translation, and the non-compositional meanings of idioms. For example, Hebrew phrases meaning ‘to fill his hand’, ‘whose hand is filled’, etc. idiomatically refer to consecrating priests; see, e.g., note to Judges

[1] They often became improperly paired with content words through fixed expressions and recurring topics.

17:5 in (van der Pool, 2003). Hilton (1981) lists sixteen citations for the idiom. However, the very common individual roots for 'fill/full' and 'hand' outweigh the distinctive combination found in the idiom.[2]

In a term-extraction approach, such as that exemplified by Melamed (2001), the collocation of 'fill' and 'hand' would attract interest because of its distribution in one of the languages, and then translations would be proposed by examining cross-language patterns (with a variety of heuristics). This almost exactly reverses our approach, wherein a pair of words attracts interest because of its cross-linguistic distribution (as calculated by a single statistic), and is then used to enrich and extend single-language studies of its members. We here pursue the word-centered approach for its simplicity and its compatibility with existing study aids, but we hope that the dialogue between Bible scholarship and computational linguistics will grow to include additional techniques and sophistication.

2. Data Preparation

The crucial data are in the CATSS parallel text (Tov, 2005), comprising the Rahlfs critical LXX aligned to the Biblia Hebraica Stuttgartensia.[3] This alignment underlies our co-occurrence data. Lemmatization data were a necessary secondary input, allowing us to re-collect co-occurrences distributed among the numerous inflected forms of both languages. We employed the CATSS LXX Morphology and the Westminster Hebrew Morphology (CATSS, 1994; Groves, 1994).

The need for parallel, lemmatized, unique, ancient data drove our text selection. Several books of LXX lacked Hebrew originals, Ben Sira and the 151st Psalm lacked lemmatizations, and Odes lacked originality, being merely a collection of poems also found elsewhere.

Three books were available in two Greek versions each. We selected Old Greek Daniel over Theodotion, since the latter postdates the New Testament. We selected Vaticanus Joshua and Judges over Alexandrinus, as do most LXX publications.[4]

We excluded the 'asterisked' passages of Job as identified by Gentry (1995),[5] accepting his argument that they were not known in Greek to the New Testament authors.

Where Hebrew tradition preserves an oral reading (Qere) different from the received written text (Ketiv), we retained both as plausible alignments to the Greek. The Masoretes, with their emphasis on the standard written text, would not have retained Qere variants at all if they did not regard them highly. Moreover, LXX supports Qere so often[6] as to suggest that (some) Qere originated before the first century C.E., when LXX fell out of Jewish favor.

Finally, where alignments between LXX and MT were debatable, we accepted the principles and therefore the conclusions of the CATSS staff. We trust the statistics to wash out oddities that result from alignment by 'formal equivalence', which is the presumption wherever plausible that the critical text of LXX is a correct and literal translation of MT. We note the reasons given by Joosten (2002) why this may not be the case—a divergent Hebrew original, divergent vocalization of identical consonants, deliberate editing, transmission error, and the accidental biases of critical reconstruction due to the effects on textual lineages of climate, politics, conquest, and happenstance. To these we might add errors in data entry, storage, heuristic repairs of these errors, and contamination of the textual data by annotations. However, we agree with him and CATSS that formal equivalence is nevertheless a necessary initial assumption to minimize subjective judgement.

We accepted CATSS's local reordering of material ('split representation') to mark uncontroversial alignments, annotating words displaced from their verse with a question mark. Where it was 'unclear whether the LXX follows the sequence of the MT or an inverted one' (Tov, 1986), CATSS followed formal equivalence but also marked the possible inversion in a back-translation; we retained both. We discarded other back-translations and questionable alternative alignments.

Data cleaning addressed versification, orthographic variation between parallel text and morphology, removal of most annotations,[7] and removal of functional words. It also addressed anomalies in the morphologies: extremely rare lemmata, very similar lemmata, and orthographic forms not unique to one lemma. Some of these were obviously incidental errors; others had to be investigated before determining whether to correct them.

A small minority of inflected words were not lemmatized because of glitches in versification, mistyped Greek diacritics, or other errors. We corrected these by inserting one or more applicable lemmata to cover 85% of the inflected type's known uses, again trusting the statistics to wash out a small amount of noise. Finally, we removed lemmata not unique within their line of the alignment, which were almost always artifacts of aligning repetitious passages rather than true double support for a word pairing.

3. Procedures

We considered each Hebrew-Greek pair within each alignment unit as a potential translation pair, then scored them with the log-likelihood statistic (Dunning, 1993), which is widely used and validated (Kageura, 1999; Melamed, 2001; Moore, 2004; Bisht et al., 2006).

We set a cutoff score for accepting a pair, following the methods of Moore (2004) so as to avoid a problem of repeated measures: Individual word pairs must meet an ex-

[2] In this project's transliteration scheme, the Hebrew roots in question are `ML)@` 'fill' (245 citations) and `YFD@` 'hand' (1479 citations), most usually translated as Greek lemmata `PI/MPLHMI` and `XEI/R` respectively.

[3] General reference on the CATSS parallel text: (Tov, 1986). Transliteration key: (Tov, 2005; Groves, 1994). We set transliterations in monospace.

[4] A list is in Danker (1993).

[5] Including several not so marked by CATSS.

[6] In the texts in our study, CATSS note 276 agreements of LXX with Qere against Ketiv, and only 169 with Ketiv against Qere.

[7] Apart from the question mark for displaced words, we retained percent and at-signs from the Westminster Hebrew Morphology to distinguish Aramaic from Hebrew proper, and occasionally part-of-speech abbreviations to distinguish between otherwise identical lemmata. All other annotations had to be characterized and removed.

tremely high standard of significance if we are to evaluate tens of thousands of them without admitting too many incorrect ones. Moore estimates noise levels in his corpus by using an exact test and concludes that the noise level drops below 1% as the log-likelihood statistic exceeds 22,[8] which we also adopted as our cutoff. For individual items, this entails a standard of significance of about $p < 0.000003$ by the X^2 approximation, or $p < 0.000005$ using Moore's direct formula, $e^{-(LLR+1.15)}$. Fortunately, suitably high scores were made possible by the extremely fine-grained and highly nonrandom alignments in CATSS. Our electronic release of the data includes all word-pairs with their log-likelihood scores, so that other users can choose a different cutoff at will.

We smoothed co-occurrence counts with Simple Good-Turing (Gale and Sampson, 1995) before computing log-likelihood, as suggested by Melamed (2001). Smoothing reduces a problem of incomplete sampling, which may mistakenly omit infrequent types and assign their probability mass elsewhere. Our smoother switched from individual calculations to curve-fitting when these were matched at $p < 0.05$.

We think it unlikely that smoothing made much difference above the cutoff (Paul Fields, personal communication); it most affects the least-supported items. However, we do consider our data a sample because they do not include Ben Sira, Psalm 151, LXX books translated from lost Hebrew originals, variations in Joshua and Judges from Alexandrinus, or translations not properly reconstructed by LXX or MT, and we undertook smoothing in an effort to treat them accordingly.

Although induction of translation lexicons in general may involve post-filtering the results (Resnik and Melamed, 1997; Melamed, 1995), our goal was not a general lexicon but a text-specific resource. This goal was better served by supplying all available information, together with statistics indicating its reliability.

A true gold standard for verifying our results would be a consultant who natively spoke both Hebrew and Greek in their Biblical form. Since there are none, we substituted second-language scholarship as a 'silver standard'. Given the large number of word pairs evaluated, we verified by sampling methods rather than exhaustively.

For a test of recall, we selected twenty word pairs from (Girdlestone, 1983), opening it at regularly spaced points and taking the nearest pair of words described as being usual.[9] We then searched for these pairs in our computed results, noting the score and rank with which the pair appeared and whether either word had a higher-ranked pairing with another.[10] This tests our approach's ability to discover word pairs that a human scholar would expect to find. Since Girdlestone covers somewhat more than 100 Hebrew words overall, we believe the sample of twenty adequately describes the relationship between his work and ours.

For a test of precision, we selected twenty pairs of words from immediately above the cutoff and noted English equivalents to the two words out of several references (Strong, 2007; Liddell et al., 1940; Brown et al., 2001; Davidson, 1900; Hurt, 2001). If these English equivalents did not clearly establish the plausibility of the word pair, we also examined the contexts in which the word pair occurred before determining its correctness. Pairs directly above the cutoff are more likely erroneous than a sample evenly distributed from the top score to the cutoff, so their precision represents a lower bound. If the thousands of word pairs above the cutoff were all evaluated, the precision would be higher. We hope that our results will eventually receive that more extensive evaluation through practical use.

4. Results and Discussion

Our main result is a file containing 38,891 Hebrew-Greek pair types, representing 256,107 pair tokens. In total it contains 5393 lemma types from Hebrew/Aramaic and 7345 from Greek. It is sorted in descending order by log-likelihood score. Its top ten pairs, given in Table 1, scored from 50,921 down to 18,640.

Hebrew	Greek	English
YHWH@	KU/RIOS	Jehovah/Lord
K.OL@	PA=S	all
B."N@	UI(O/S	son
LO)@	OU)	not
)MR@	EI)=PON	speak
MELEK:@	BASILEU/S	king
(&H@	POIE/W	do/make
):ELOHIYM@	QEO/S	god
)EREC@	GH=	land
YOWM@	H(ME/RA	day

Table 1: Top ten overall results.

A similar, smaller file was prepared only from the Aramaic data. It contains 1474 pair types representing 3245 tokens, 551 lemma types from Aramaic and 807 from Greek. In the main results, the Aramaic words appear as unusual synonyms of their better-attested Hebrew cognates, which bury them. This file was not subjected to precision/recall tests independently, but its score distribution strongly resembles the Hebrew results in miniature, and its top results, given in Table 2, are similar except for some topical bias.

The Moore (2004) cutoff for the 1% noise level admits 7292 items from the overall results (18.7% of the total), and 256 items from the Aramaic-only (17.4%).

Both files are available in full online, as are the co-occurrence counts and the underlying aligned, lemmatized content words. We also provide a concordancing program, called cite-puller, which provides citations for cross-language pairs as well as for single lemmata. This makes it easy to explore the textual evidence behind the log-likelihood score. All are available at http://www.

[8] He systematically omits a factor of 2 and so states this value as 11.

[9] In one case the only Greek nearby was not so described. More on this below.

[10] Because our Hebrew lemmatization usually omits vowels and Girdlestone does not, in practice this meant searching from the top of the results for the Greek form. We accepted the Hebrew lemma as matching if it was a subsequence of the full spelling in Girdlestone.

Aramaic	Greek	English
`MELEK:%`	`BASILEU/S`	king
`):ELFH.%`	`QEO/S`	god
`K.OL%`	`PA=S`	all
`B.AYIT%`	`OI)=KOS`	house
`MAL:K.W.%`	`BASILEI/A`	kingdom
`LF)%`	`OU)`	not
`)MR%`	`EI)=PON`	speak
`$:MAYIN%`	`OU)RANO/S`	skies/heaven
`B.NH%`	`OI)KODOME/W`	build
`QWM%`	`I(/STHMI`	stand

Table 2: Top ten Aramaic results.

`ttt.org/mtlxx/index.html` for academic and personal study.[11]

4.1. Recall

Of twenty pairs selected from Girdlestone, fourteen had the top score for both their Hebrew and their Greek word. These pairs are given in Table 3.

Hebrew	Greek	English
`):ELOHIYM@`	`QEO/S`	god
`)FDFM@`	`A)/NQRWPOS`	human being
`RW.XA@`	`PNEU=MA`	spirit
`X+)@`	`A(MARTA/NW`	sin
`P.DH@`	`LUTRO/W`	deliverance
`N&)@`	`AI)/RW`	bear, carry
`NZH@`	`R(AI/NW`	sprinkle
`M$X@`	`XRI/W`	anoint
`$X+@`	`SFA/ZW`	slay
`B.:RIYT@`	`DIAQH/KH`	covenant
`H"YKFL@`	`NAO/S`	temple
`K.OH"N@`	`I(EREU/S`	priest
`)BD@`	`A)PO_O)LLU/W`	destroy
`&F+FN@`	`DIA/BOLOS`	accuser

Table 3: Perfect agreements with Girdlestone.

Three more were top-scoring for Hebrew, but the Greek word scored higher in one or two other pairs: `CEDEQ@ – DIKAIOSU/NH` 'righteousness' and `$IQ.W.C@ – BDE/LUGMA` 'abomination' were the second matches for their Greek members, and `YXL@ – E)LPI/ZW` 'hope' was the third. In the case of 'righteousness' and 'hope' the higher-ranked words are nearby in Girdlestone, so only an accident of sample selection demotes them from the previous list. The higher match for 'abomination', `$IQ.W.C@`, is a true synonym, whose absence from Girdlestone we construe as his error.

The lowest log-likelihood among these seventeen pairs was 246, suggesting that the noise cutoff of 22 easily captures not only the pairs of primary importance, but important secondary associations as well. One more sampled pair scored 47, modestly above the cutoff: `$AD.AY@ – QEO/S` pairs 'Shaddai' (a rare Hebrew divine title of contested meaning) with the generic Greek word for 'god' (which has several better matches in Hebrew). We believe its low score appropriate to the poor match it represents. Girdlestone lists several other Hebrew equivalents to 'God' (nine outrank 'Shaddai' in our statistics), but he does not mention translating 'Shaddai' as `PANTOKRA/TWR` (its top match, scoring 163), which we regard as another error.

The final two pairs in the sample are definite mismatches in Girdlestone: First, `Y+B@ – A)GAQO/S` pairs a stative verb 'be good' with an adjective 'good', despite other words better matched to them in semantics, grammar, and frequency. Second, `+AP@ – OI)KI/A` 'household' was selected under duress; one of the evenly spaced sampling points fell amid a lengthy discussion of genealogical terminology almost totally devoid of Greek, and although Girdlestone mentions this pair in reviewing another scholar's views, he neither claims it himself nor calls it usual.[12] These two pairs scored 2 and 1 respectively, far below the cutoff.

Based on 18/20 successes, we might claim 90% recall against the 'silver standard'. We believe it more appropriate to say that our project *outperforms* Girdlestone, in light of the weaknesses this test exposed.

4.2. Precision

We translated to English the twenty pairs scoring just above the cutoff, via reference works cited above. Their log-likelihood scores range from 22.001 to 22.055. Twelve are plainly unproblematic. Three more turn out to be sensible on inspecting the contexts where the words co-occur. Four more reflect textual correlates but not translation pairs, often where the text is in dispute. One is frankly wrong, an accidental consequence of multi-word Hebrew numbers.

If a pure translation dictionary is desired, this means precision is above 75% (we expect higher-ranked pairs to be more precise). However, we desire to find textual correlates that are not translations, which adds four more pairs and raises lower-bound precision to 95%. The results are shown in Table 4.

Some of these are more definitive translation pairs than others, of course. We expect the user to take caution from their comparatively low scores, and to read them where possible in conjunction with other, higher-ranked pairs for their constituent words. On these terms they are informative and useful.

4.3. Discussion

Recall and precision samples show that our method produces sound pairs of corresponding words: They easily cover a variety of important lexis, without radical departures from bilingual scholars' views. Even those nearest the cutoff include some excellent translation pairs. Where the translation is less straightforward, they signpost important textual difficulties.

Words representing numbers are problematic in our results. `$FLO$` 'three' and `TRISKAI/DEKA` 'thirteen' did not cooccur by chance, but because of the Hebrew grammar of numbers. This is a caution to apply these results only with an eye to their origin. On the other hand, the pair's low

[11] Thanks to Alan K. Melby for hosting these.

[12] Our statistics instead suggested `B.AYIT@`, a much more frequent Hebrew word, with a score of 1831.

Hebrew	English from Hebrew	English from Greek	Greek	Remarks
L:BFNFH@	the moon	the moon	SELH/NH	
RAQ@	lean, thin	peeled, threshed, small, lean, thin, refined, poor	LEPTO/S	
N(M@	delight, splendor, be agreeable, grace	season, make pleasant, delight, gratify	H(DU/NW	
K.HH@	be weak, despondent, (eye) grow dim	make blind	E)K_TUFLO/W	
(FR:LFH@	foreskin	hardheartedness	SKLHROKARDI/A	(1)
NQB@	perforate, bore, libel, blaspheme	pierce, bore	TRUPA/W	
P.(L@	do, make, practice, work	complete, accomplish	E)K_E)RGA/ZOMAI	
XRD@	fear, hasten, discomfit	drive away, agitate, strut	SOBE/W	(2)
XRD@	fear, hasten, discomfit	scare away, keep off, be off	A)PO_SOBE/W	(2)
P.AR:T.:MIYM@	nobles, princes	splendid, noble	E)/NDOCOS	
$AL:WFH@	security, abundance	suddenly	E)CA/PINA	(3)
YKX@	be right, argue, decide, convict	refute, convict, try	DIA_E)LE/GXW	
K.FT"P@	shoulder, corner, side-piece	shoulder, tunic, ephod, door leaf	E)PWMI/S	
&YM@	place, set, determine, leave, order	stand upon, be at hand, set over, establish	E)PI_I(/STHMI	(4)
(AM@	people, tribe	home, family	OI)=KOS	
&AR@	captain	commander of ten	DEKA/DARXOS	
FLO@	three	thirteen	TRISKAI/DEKA	(5)
YCR@	press, narrow, distressed, potter, fashion	smelting furnace	XWNEUTH/RION	(6)
NP$@	breathe, refresh	breathe, chill	YU/XW	
HLL@	boast, celebrate, commend, foolish, praise, renowned, shine	acknowledge, confess, profess, promise	E)K_O(MOLOGE/W	(7)

(1) Deut 10:16, Jer 4:4. Consistent MT/LXX difference.
(2) Deut 28:26, Jer 7:33. Usage makes clear where words' senses overlap.
(3) Dan 11: 21, 24; LXX text disputed. van der Pool (2003) instead has EU)QENI/A 'prosperity'.
(4) Comparably broad verbs.
(5) Due to multi-word Hebrew expressions for numbers.
(6) Zech 11:13. Consistent MT/LXX difference; Ellinger & Rudolph (1997) suggest)OWCFR 'treasury' from Syriac.
(7) 1Chr 23:30, 2Chr 5:13, 2Chr 31:2. Each word has stronger matches, but they are used inconsistently in Chronicles.

Table 4: Near-cutoff results from aligned extraction.

score is also a caution, and the user who searches for better matches will soon discover how FLO associates with *various* threes. The supposition that perhaps it is 'three' and not 'thirteen' should follow in due course.

YCR@ 'potter' mismatched with XWNEUTH/RION 'furnace' illustrates signposting of textual difficulty. Extant information cannot determine whether Zech 11:13 originally concerned either or neither. Joosten (2002) argues from MT/LXX divergences that we cannot blindly adopt just any extant Hebrew-Greek pairing as if it were an exact gloss, and we find, particularly with infrequent words, that this is still the case.

However, if our statistics cannot relieve us of caution, they can at least relieve us of doubt. Higher scores indicate more stereotyped and widespread translations where we may have 'greater confidence that the translators' knowledge of Hebrew is operating' (Joosten, 2002), where the word-study movement's traditional lexicographic approach is sufficient. Lower scores pertain to unusual translations motivated by unusual contexts (Tov, 1986), which call for a more philological and text-critical study.

Strictly speaking, the word pairs scored by this project tell us only about Hebrew-to-Greek translation.[13] But it is the ideal of translation to be reversible, and LXX emphasizes this ideal by its literal approach. Moreover, log-likelihood is symmetrical; the scores would be the same if the Greek had been original, and the Hebrew were the product of translation in the opposite direction. Finally, because the Septuagint was broadly used in Judaism for three centuries, the words of Biblical Hebrew and those of Septuagint-like Greek (though not necessarily Koine Greek at large) partly mingled their meanings (Dodd, 1935). Therefore, in the case of New Testament Greek, there is at least linguistic and literary justification, if not statistical, for applying our results in the Greek-to-Hebrew direction.

5. Conclusion and Future Work

The present results might easily be refined by improving the source data. The Hebrew lemmatization data has been revised, though the revision was not available to us. Ben Sira and the 151st Psalm could be lemmatized by analogy to data in hand. Joshua and Judges could be prepared in a combined Alexandrinus/Vaticanus version, including their differences on equal grounds just as we did Ketiv/Qere. Several fixed phrases of function words have well-established non-compositional meanings, which might be coded as a lemma before the function words themselves are eliminated. Finally, the noise of number words might be eliminated.

The study could be expanded to a publishable reference work, intermediate between Girdlestone's theological essays and the dense data of Hatch-Redpath. This would require setting a less-arbitrary cutoff, designing the presentation of the information, associating the lemmata and pairs with additional information for cross-referencing, and obtaining permission from copyright holders of the source data to commercialize a derivative work. The information design for such a reference has already been sketched.

A related approach might preprocess each language to identify interesting collocations. Multi-word units might be selected as in Melamed (2001), or the co-occurrence score alone might be used to identify distinctive, consistent phrasal translations after exhaustively annotating all local word combinations with a Sparse Binary Polynomial Hash (Yerazunis, 2003).

A study of function words would require a very different translation model. At a guess, it would have to take account of syntactic dependencies, immediate lexical context, and topic, as a human translator does. Topic might be measured by distance in-text to several key topical words, modifying the centrality measures of Kwon (2007). It remains unclear, a priori, whether such a study would reveal anything of interpretive interest.

This work demonstrates the power of a data-centric approach to overcome a known difficulty in Bible studies. We hope that it will not only prove useful for studies of intertestamental rhetorical/interpretive relations, but also suggest further investigations in computational philology.

[13] Thanks to Paul Fields (personal communication) for pointing this out.

6. References

Raj Kishor Bisht, H. S. Dhami, and Neeraj Tiwari. 2006. An evaluation of different statistical techniques of collocation extraction using a probability measure to word combinations. *Journal of Quantitative Linguistics*, 13:161–175.

Francis Brown, S. R. Driver, and Charles A. Briggs. 2001. *The Brown-Driver-Briggs Hebrew and English lexicon: With an appendix containing the Biblical Aramaic: Coded with the numbering system from Strong's exhaustive concordance of the Bible*. Hendrickson, Peabody, Massachusetts.

CATSS. 1994. Morphologically analyzed Septuagint. `http://ccat.sas.upenn.edu/gopher/text/religion/biblical/lxxmorph/`.

Peter Cotterell and Max Turner. 1989. *Linguistics and biblical interpretation*. SPCK, London.

Frederick W. Danker. 1993. *Multipurpose tools for Bible study*. Fortress Press, Minneapolis, Minnesota, revised edition.

Benjamin Davidson. 1900. *The analytical Hebrew and Chaldee lexicon*. James Pott, New York.

C. H. Dodd. 1935. *The Bible and the Greeks*. Hodder and Stoughton, London.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–64.

Karl Elliger and Wilhelm Rudolph. 1997. *Biblia Hebraica Stuttgartensia*. Deutsche Bibelgesellschaft, Stuttgart, Germany, 5 edition.

William A. Gale and Geoffrey Sampson. 1995. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2:217–237.

Peter John Gentry. 1995. *The asterisked materials in the Greek Job*, volume 38 of *Septuagint and Cognate Studies*. Scholars Press, Atlanta, Georgia.

Robert Baker Girdlestone. 1983. *Synonyms of the Old Testament: Their bearing on Christian doctrine*. Baker, Grand Rapids, Michigan, 3rd edition edition.

Alan Groves. 1994. Westminster Hebrew morphological database. More recent versions are known as the Groves-Wheeler Hebrew Morphology. Westminster Theological Seminary, Glenside, Pennsylvania.

Edwin Hatch and Henry A. Redpath. 1998. *A Concordance to the Septuagint and the other Greek versions of the Old Testament (including the Apocryphal books)*. Baker, Grand Rapids, Michigan, 2nd edition.

Lynn M. Hilton. 1981. The hand as a cup in ancient temple worship. Paper given at the Thirtieth Annual Symposium on the Archaeology of the Scriptures, held at BYU on 26 September 1981. Available from `http://www.mormonmonastery.org/230-the-hand-as-a-cup-in-ancient-temple-worship/`.

John Hurt. 2001. Strong's greek dictionary.

Full text, transliterated. Available from http://www.sacrednamebible.com/kjvstrongs/STRINDEX.htm.

Jan Joosten. 2002. Biblical Hebrew as mirrored in the Septuagint: The question of influence from spoken Hebrew. *Textus*, 21:1–20.

Kyo Kageura. 1999. Bigram statistics revisited: A comparative examination of some statistical measures in morphological analysis of Japanese kanji sequences. *Journal of Quantitative Linguistics*, 6:149–166.

Kyounghee Kwon. 2007. Assessing semantic equivalence for shared understanding of international documents: An application of semantic network analysis to multilingual translations of the Universal Declaration of Human Rights. Master's thesis, State University of New York, Buffalo, New York.

Henry George Liddell, Robert Scott, and Sir Henry Stuart Jones. 1940. *A Greek-English lexicon*. Clarendon, Oxford.

I. Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 184–198.

I. Dan Melamed. 2001. *Empirical methods for exploiting parallel texts*. MIT Press, Cambridge, Massachusetts.

Robert C. Moore. 2004. On log-likelihood ratios and the significance of rare events. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 333–340, Barcelona. Association for Computational Linguistics.

Takamitsu Muraoka. 1988. *Hebrew-Aramaic index to the Septuagint, keyed to the Hatch-Redpath concordance*. Baker, Grand Rapids, Michigan.

Philip Resnik and I. Dan Melamed. 1997. Semi-automatic acquisition of domain-specific translation lexicons. In *Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, D.C.*, pages 340–347, Stroudsburg, Pennsylvania. Association for Computational Linguistics.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a parallel corpus: Annotating the "Book of 200 tongues". *Computers and the Humanities*, 33:129–153.

James Strong. 2007. *Strong's exhaustive concordance to the Bible: Updated edition with CD*. Hendrickson, Peabody, Massachusetts.

Emanuel Tov. 1986. *A computerized data base for Septuagint studies: The Parallel Aligned Text of the Greek and Hebrew Bible*, volume 2 of *Computer Assisted Tools for Septuagint Studies*. Journal of Northwest Semitic Languages, Stellenbosch, South Africa.

Emanuel Tov. 2005. The parallel aligned Hebrew-Aramaic and Greek texts of Jewish scripture. http://ccat.sas.upenn.edu/gopher/text/religion/biblical/parallel/.

Charles L. van der Pool. 2003. *The apostolic Bible polyglot*. Apostolic Press, Newport, Oregon, pdf edition.

William S. Yerazunis. 2003. Sparse Binary Polynomial Hashing and the CRM114 Discriminator. Version of 20 January 2003. Available from http://crm114.sourceforge.net/docs/CRM114_paper.html.

Authorship Classification of two Old Arabic Religious Books Based on a Hierarchical Clustering

Halim Sayoud
Faculty of Electronics and Informatics
USTHB University, Algiers, Algeria
www.usthb.dz
halim.sayoud@uni.de

Abstract

Authorship classification consists in assigning classes to a set of different texts, where each class should represent one author. In this investigation, the author presents a stylometric research work consisting in an automatic authorship classification of eight different text segments corresponding to four text segments of the Quran (The holy words and statements of God in the Islamic religion) and four other text segments of the Hadith (statements said by the prophet Muhammad).

Experiments of authorship attribution are made on these two old religious books by employing a hierarchical clustering and several types of original features. The sizes of the segments are more or less in the same range.

The results of this investigation shed light on an old religious enigma, which has not been solved for fifteen hundred years: results show that the two books should have two different authors or at least two different writing styles.

Keywords: Authorship classification, Authorship attribution, Origin of old books, Quran, Prophet’s Statements, Hierarchical clustering.

1. Introduction

Authors have different ways of speaking and writing (Corney, 2003), and there exists a rich history of linguistic and stylistic investigation into authorship classification (Holmes 1998). In recent years, practical applications of authorship attribution have grown in areas such as intelligence, criminal law, civil law and computer security. This activity is part of a broader growth within computer science of identification technologies, including biometrics, cryptographic signatures, intrusion detection systems, and others (Madigan, 2005).

In the present paper, we deal with a religious enigma, which has not been solved for fifteen hundred years (Sayoud, 2010). In fact, many critics of Islam throughout their continuous searching to find a human source for the Holy Quran do exist; their imagination conducted them to the idea that the Holy Quran was an invention of the prophet Muhammad (Al-Shreef web).

Several theologians, over time, tried to prove that this assumption was false. They were relatively logical and clever, but their proofs were not so convincing for many people, due to a lack in the scientific rigor.

Similarly, for the Christian religion, there exist several disputes about the origin of some texts of the Bible. Such disputes are very difficult to solve due to the delicacy of the problem, the religious sensitivity and because the texts were written a long time ago.

One of the purposes of stylometry is authorship attribution, which is the determination of the author of a particular piece of text for which there is some dispute about its writer (Mills, 2003).

Hence, it can be seen why Holmes (Mills, 2003) pinpointed that the area of stylistic analysis is the main contribution of statistics to religious studies. For example, early in the nineteenth century, Schleiermacher disputed the authorship of the Pauline Pastoral Epistle 1 Timothy (Mills, 2003). As a result, other German speaking theologians, namely, F.C. Baur and H.J. Holtzmann, initiated similar studies of New Testament books (Mills, 2003).

In such problems, it is crucial to use rigorous scientific tools and it is important to interpret them very carefully.

In this paper, we try to make an authorship discrimination experiment (Jiexun, 2006) between the Quran and some Prophet’s statements, which is based on a hierarchical clustering, in order to show that the Quran was not written by the Prophet Muhammad, if the results of this technique confirm that supposition (Al-Shreef web).

The manuscript is organized as follows: Section 2 gives a description of the two books to be compared. Section 3 discusses the following philosophic problem: *Was the Quran invented by the prophet Muhammad?* Section 4 describes the different experiments of authorship classification. Section 5 displays the results of the hierarchical clustering and an overall discussion is given at the end of the manuscript.

2. Description of the two Books

A brief description of the two investigated books (*Quran and Hadith*) is provided in the following subsections.

2.1 The Quran

The Quran (in Arabic: القرآن al-qur’ān, literally "the recitation"; also sometimes transliterated as Qur’ān, Koran, Alcoran or Al-Qur’ān (Wiki1, 2012) (Nasr, 2007)) is the central religious text of Islam. Muslims believe the Quran to be the book of divine guidance and direction for mankind (Ibrahim, 1996) (that has been written by God), and consider this Arabic book to be the final revelation of God. Islam holds that the Quran was written by Allah (ie. God) and transmitted to Muhammad by the angel Gibraele (Gabriel) over a period of 23 years. The beginning of Quran apparition was in the year 610 (after the Christ birth).

2.2 The Hadith

Hadith (in Arabic: الحديث, transliteration: al-ḥadīth (Wiki2, 2012) (Islahi, 1989)) is the oral statements and words said by the Islamic prophet Muhammad (Pbuh). Hadith collections are regarded as important tools for determining the Sunnah, or Muslim way of life, by all traditional schools of jurisprudence. In Islamic terminology, the term hadith refers to reports about the statements or actions of the Islamic prophet Muhammad, or about his tacit approval of something said or done in his presence (Wiki2, 2012) (Islahi, 1989). The text of the Hadith (matn) would most often come in the form of a speech, injunction, proverb, aphorism or brief dialogue of the Prophet whose sense might apply to a range of new contexts. The Hadith was recorded from the Prophet for a period of 23 years between 610 and 633 (after the Christ birth).

3. Was the Quran invented by the prophet Muhammad?

Muslims believe that Muhammad was only the narrator who recited the sentences of the Quran as written by Allah (*God*), but not the author. See what Allah (*God*) says in the Quran book: « O Messenger (*Muhammad*)! transmit (*the Message*) which has been sent down to you from your Lord. And if you do not, then you have not conveyed his Message. Allah will protect you from people. Allah do not guide the people who disbelieve » (5:67).

Some critics of Islam throughout the continuous search to find a human source for the Holy Quran do exist; such assumptions suppose that the Holy Quran is an invention of the prophet Muhammad (Al-Shreef web).

For a long time, different religious scientists presented strong context-based demonstrations showing that this assumption is impossible.

The purpose of our research work is to conduct a text-mining based investigation (*ie. authorship classification*) in order to see if the two concerned books have the same or different authors (Mills, 2003) (Tambouratzis 2000) (Tambouratzis, 2003), regardless of the literal style or context.

4. Experiments of automatic authorship classification

4.1 Summary on the dimension of the two books

This subsection summarizes the size of the two books in terms of words, tokens, pages, etc. The different statistical characteristics of the two books are summarized as follows:

- Number of Tokens in the Quran= 87341
- Number of Tokens in the Hadith= 23068
- Number of different words in the Quran= 13473.
- Number of different words in the Hadith= 6225.
- Average Number of A4 pages in the Quran= 315 pages.
- Average Number of A4 pages in the Bukhari Hadith= 87 pages.
- Ratio of the Number of Quran Tokens / Number of Hadith Tokens = 3.79
- Ratio of the Number of Quran Lines / Number of Hadith Lines of Bkh = 3.61
- Ratio of the Number of different Quran words / Number of different Hadith words = 2.16
- Ratio of the Number of Quran A4 Pages / Number of Hadith A4 Pages = 3.62

The two books seem relatively consistent since the average number of A4 pages is 315 for the Quran book and 87 for the Hadith book. However, the 2 books do not have the same size; that is why we should proceed with our investigation with care. In this investigation we chose 4 different segments from the Quran and 4 other segments from the Hadith, in order to handle texts with more or less a same size range.

4.2 Experiments based on a segmental analysis

These experiments analyses the two books in a segmental form: four different segments of texts are extracted from every book and the different segments are analyzed in a purpose of authorship verification. It concerns five experiments: an experiment using discriminative words, a word length frequency based analysis, an experiment using the COST parameter, an investigation on discriminative characters and an experiment based on vocabulary similarities.

In these experiments, the different segments are chosen as follows: one segment is extracted from the beginning of the book, another one from the end and the two other segments are extracted from the middle area of the book. A segment size is about 10 standard A4 pages (between 5000 and 7700 tokens) and all the segments are distinct and separated (*without intersection*). These segments are denoted Q1 (*or Quran 1*), Q2 (*or Quran 2*), Q3 (*or Quran 3*), Q4 (*or Quran 4*), H1 (or Hadith 1), H2 (*or Hadith 2*), H3 (*or Hadith 3*) and H4 (*or Hadith 4*). In these experiments, the different segments are chosen as follows: one segment is extracted from the beginning of the book, another one from the end and the two other segments are extracted from the middle area of the book. Finally, these eight texts segments are more or less comparable in size.

4.2.1 Discriminative words

This first experiment investigates the use of some words that are very commonly employed in only one of the book. In practice, we remarked that the following words: الذين (*in English: THOSE or WHO in a plural form*) and الأرض (*in English: EARTH*) are very commonly used in the four Quran segments; whereas, in the Hadith segments, these words are rarely used, as we can see in the following table.

Word	Frequency (%)							
	Q1	Q2	Q3	Q4	H1	H2	H3	H4
الذين	1.35	1.02	1.12	0.75	0.11	0.03	0.02	0.08
الأرض	0.34	0.63	0.59	0.42	0.23	0.13	0.18	0.15

Table 1: Some discriminative words and their frequencies.

For الذين the frequency of occurrence is about 1% in the Quran segments, but it is between 0.02% and 0.11% in the Hadith segments (*namely almost the 1/10th of the Quran frequency*).
For الأرض the frequency of occurrence is about 0.5% in the Quran segments, but it is between 0.13% and 0.23% in the Hadith segments (*namely about the half*).

4.2.2 Word length frequency based analysis
The second experiment is an investigation on the word length frequency. The following figure (figure 1), which contains the different smoothed curves representing the « word length frequency » versus the « word length », shows the following two important points:

- The Hadith curves have more or less a gaussian shape that is pretty smooth; whereas the Quran curves seem to be less Gaussian and present some oscillations (distortions).
- The Hadith curves are easily distinguishable from the Quran ones, particularly for the lengths 1,3, 4 and 8: for the lengths 1 and 8, Quran possesses higher frequencies, whereas for the lengths 3 and 4, Hadith possesses higher frequencies.

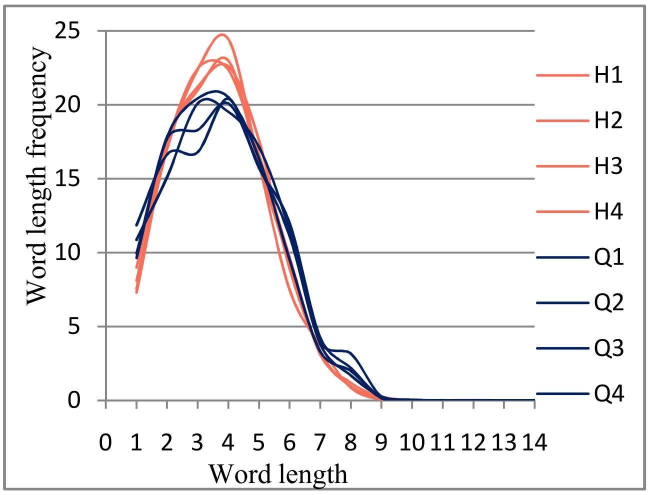


Figure 1: Word length frequency (smoothed lines).

4.2.3 A new parameter called COST

The third experiment concerns a new parameter that we called COST and which appears non-null only in the Holy Quran, as we can see in the following table 2.
The COST parameter is a cumulative distance measuring the similarity between the ending of one sentence and the ending of the next and the previous one in the text. It gives an estimation measure on the poetic form of a text. In fact, when poets write a series of poems, they make a termination similarity between the neighboring sentences of the poem, such as a same final syllable or letter. That is, the COST parameter estimates the similarity ratio between successive sentences in term of ending syllables.

The following table 2 shows the average COST values of the 8 different segments.

	Q1	Q2	Q3	Q4	H1	H2	H3	H4
COST_{avr}	2.2	2.6	2.6	2.38	0.46	0.47	0.43	0.47

Table 2: Average COST values for the different segments.

By observing the above table, we notice that the average value of the COST is practically constant for all the Quran segments: it is about 2.2 at the beginning and end of the Quran and it is about 2.6 in the area of the middle.
Similarly, this parameter appears constant for all the Hadith segments: it is about 0.46.
In addition, we notice that the mean values of the COST for Quran and Hadith are very different.

4.2.4 Discriminative characters

The fourth experiment investigates the use of some characters that are very commonly used in only one of the books.
In reality, we limited our investigation to one of the most interesting character, which seems to be very discriminative between the two books: it is the character "و" , which is a consonant and vowel in a same time (*in English, it is equivalent to the consonant W when used as consonant; or the vowel U when used as vowel*).
Furthermore, this character is important because it also represents the preposition AND (*in English*), which is widely used in Arabic.
So, by observing the table below, we notice that this character has a frequency of about 7% in all Quran segments and a frequency of about 5% in all Hadith segments.

Seg.	Q1	Q2	Q3	Q4	H1	H2	H3	H4
Freq. of “و“	7.73	7.11	6.91	7.04	5.19	5.45	4.72	5.33

Table 3: frequency of the character و in %

This difference in the frequencies implies two different ways of using the character و (which is also a conjunction).

4.2.5 Vocabulary based similarity

The fifth experiment makes an estimation of the similarity between the vocabularies (*words*) of the two books.
So, in this investigation we propose a new vocabulary similarity measure that we called VSM (*ie. Vocabulary Similarity Measure*), which is defined as follows:

VSM (text1, text2) = (number of common words between the 2 texts) / (size(text1) . size(text2))$^{1/2}$ (1)

Typically, in case of 2 identical texts, this similarity measure will have a value of 1 (*ie. 100%*). That is, the much higher this measure is, the much similar (*in vocabulary*) are the two texts.

The different inter-measures of similarity are represented in the following matrix (*similarity matrix*), which is displayed in table 4.

VSM in %	H1	H2	H3	H4	Q1	Q2	Q3	Q4
H1	100	32.9	31.4	28.2	20.9	19.9	19.4	19.9
H2	32.9	100	31.4	29.2	20.8	20.0	18.6	19.5
H3	31.4	31.4	100	29.2	19.8	19.9	18.9	19.0
H4	28.2	29.2	29.2	100	19.9	18.7	18.6	18.8
Q1	20.9	20.8	19.8	19.9	100	29.7	29.6	24.5
Q2	19.9	20.0	19.9	18.7	29.7	100	34.9	25.2
Q3	19.4	18.6	18.9	18.6	29.6	34.9	100	27.1
Q4	19.9	19.5	19.0	18.8	24.5	25.2	27.1	100

Table 4: Similarity matrix representing the different VSM similarities between segments.

We notice that all the diagonal elements are equal to 100%. We do remark also that all the Q-Q similarities and H-H similarities are high, relatively to Q-H or H-Q ones (*Q stands for a Quran segment and H stands for a Hadith segment*). This means that the 4 segments of the Quran have a great similarity in vocabulary and the 4 segments of the Hadith have a great similarity in vocabulary, too. On the other hand it implies a low similarity between the vocabulary styles of the two different books. This deduction can easily be made from the following simplified table, which represents the mean similarity measure between one segment and all the segments of a same book.
Table 5 gives the mean similarity according to Quran or Hadith for each segment X ($X=Q_i$ *or* $X=H_{i,}$ $i=1..4$), which can be expressed as the average of all the similarities between that segment (*X*) and all the other segments (*excluding segment X*) of a same type of all segments. This table is displayed in order to see if a segment is much more similar to the Quran family or to Hadith family.

	Mean Similarity with H segments	Mean Similarity with Q segments
H1	30.85	20.01
H2	31.16	19.73
H3	30.66	19.38
H4	28.87	18.99
Q1	20.37	27.92
Q2	19.60	29.94
Q3	18.87	30.51
Q4	19.27	25.60

Table 5: Mean VSM similarity in % between one segment and the different segments of a same book.

Similarly, we remark that the intra-similarities (*within a same book*) are high: between 26% and 31%; and that the inter-similarities (*segments from different books*) are relatively low: not exceeding 20%.

4.2.6 Exploitation of these results

In order to exploit all the previous results, a hierarchical clustering (Miro 2006) is employed by fusing all the numerical scores into a global feature vector for every segment (*there are 8 different segments*).
The results of the hierarchical clustering are described in section 5.

5. Hierarchical clustering based classification

In this section we will use a hierarchical clustering (Miro, 2006) (with an average linkage) and try to separate the 8 segments into different appropriate clusters (hopefully 2 clusters), thanks to the different features (as described previously), which are denoted in this investigation as follows:

- F1= frequency of the word (الذين)
- F2= frequency of the word (الأرض)
- F3= frequency of words with a length of 1 character
- F4= frequency of words with a length of 2 characters
- F5= frequency of words with a length of 3 characters
- F6= frequency of words with a length of 4 characters
- F7= frequency of words with a length of 5 characters
- F8= frequency of words with a length of 6 characters
- F9= frequency of words with a length of 7 characters
- F10= frequency of words with a length of 8 characters
- F11= frequency of words with a length of 9 characters
- F12= frequency of the character (و)
- F13= COST value
- F14= Average Vocabulary Similarity Measure with regards to the Quran
- F15= Average Vocabulary Similarity Measure with regards to the Hadith

The employed distance is the cosine distance and all the previous result scores are fused into a global feature vector for each segment. These feature vectors are used as input vectors for the hierarchical clustering.
The result of this hierarchical clustering is given by the following dendrogram (see figure 2), which illustrates the different possibilities of clustering with their corresponding distances in a graphical way.

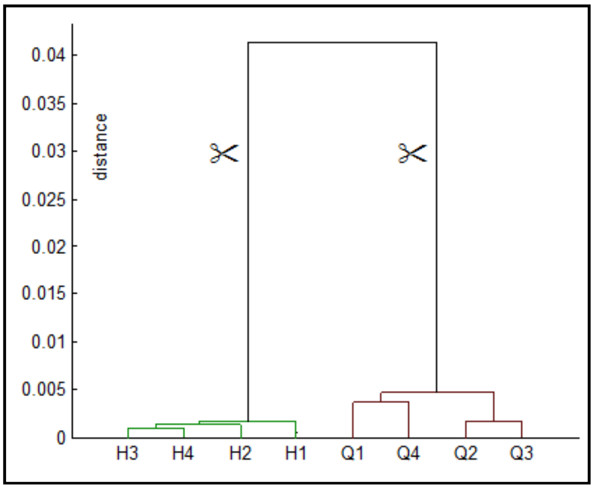


Figure 2: Dendrogram of the different clusters with the corresponding distances. Note that the last cluster (big one) is inconsistent, which involves two main classes.

The inconsistencies of the different clusters are respectively (*from the first cluster to the last one*): 0, 0.7, 0, 0.7, 0, 0.85 and 1.15.

As we can see, the last cluster has an inconsistency parameter greater than 1 (*inconsistency of 1.15*), while all the other clusters do not exceed 0.85.

Moreover, even by observing the dendrogram of figure 2, we can easily notice that the hierarchical clustering has revealed two main classes: one class grouping the 4 Hadith segments (*in the left side of the dendrogram*) and a second class grouping the 4 Quran segments (*in the right side of the dendrogram*).

This new result involves two important conclusions:

- First, the two books Q and H should have different authors;
- Second, the 4 segments of each book seem to have a same author (*the presumed author of the book*).

6. Discussion

In this research work, we have made an investigation of authorship discrimination (Tambouratzis 2000) (Tambouratzis, 2003) between two old Arabic religious books: the Quran and Bukhari Hadith.

For that purpose, 15 different features, which have been extracted from 5 different experiments, are employed and tested on 8 different text segments (*corresponding to 4 Quran segments and 4 Hadith segments*).

Thereafter, a hierarchical clustering has been applied to group all the text segments into an optimal number of classes, where each class should represent one author.

Results of this experiment led to three main conclusions:

- First, there are probably two different classes, which correspond to two different authors;
- Second, the two investigated books appear to have different authors or at least two different writing styles;
- Third, all the segments that have been extracted from a unique book (*from the Quran only or from the Hadith only*) should have a same author.

These results show that the Quran could not be written by the Prophet Muhammad and that it should belong to a unique author too. Muslims believe that it is written by Allah (*God*) and sent to his messenger (*the prophet Muhammad*). We will not extend our research work into an etymological point of view: this is not the main topic of this work, but we think that it may be interesting to mention this point.

7. Acknowledgements

The author wishes to thank Dr George Tambouratzis and Dr Marina Vassiliou from the ILSP institute of Athens and Dr Efstathios Stamatatos from the Aegean University of Samos for their great helps and advices. He also wishes to thank Dr Patrick Juola from Duquesne University for his pertinent comments regarding our research works. I do not forget to express many thanks to Pr Michael Oakes from Sunderland University for his constructive comments and generous help. Finally, in order to enhance the quality of this research work, the author welcomes all the comments and suggestions of the readers regarding this topic.

8. References

Al-Shreef. A. Is the Holy Quran Muhammad's invention ? http://www.quran-m.com/firas/en1/index.php?option=com_content&view=article&id=294:is-the-holy-quran-muhammads-invention-&catid=51:prophetical&Itemid=105

Clement R., Sharp. D. (2003). Ngram and Bayesian Classification of Documents for Topic and Authorship. Literary and Linguistic Computing, 18(4), pp. 423–447, 2003.

Corney. M. W. (2003). Analysing E-Mail Text Authorship for Forensic Purposes. Master Thesis, Queensland University of Technology, Australia, 2003.

Holmes. D. I. (1998). The Evolution of Stylometry in Humanities Scholarship . Literary and Linguistic Computing, Vol. 13, No3, pp. 111-117.

Ibrahim. I. A. (1996). A brief illustrated guide to understanding Islam. Library of Congress, Catalog Card Number: 97-67654, Published by Darussalam, Publishers and Distributors, Houston, Texas, USA. Web version: http://www.islam-guide.com/contents-wide.htm, ISBN: 9960-34-011-2.

Li, J., Zheng, R., and Chen, H. (2006). From fingerprint to writeprint. Communications of the ACM, vol 49, No 4, April 2006, pp. 76-82.

Juola. P. (2009). JGAAP: A System for Comparative Evaluation of Authorship Attribution. Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science, Vol. 1, No 1, 2009.

Madigan D., Genkin, A. Lewis, D. Argamon, S. Fradkin, D. and Ye. L. (2005). Author identification on the large scale. In Joint Annual Meeting of the Interface and the Classification Society of North America (CSNA), 2005.

Mills. D. E. (2003). Authorship Attribution Applied to the Bible. Master thesis, Graduate Faculty of Texas, Tech University, 2003.

Miro. X. A. (2006). Robust Speaker Diarization for meetings, PhD thesis, Universitat Polit`ecnica de Catalunya, Barcelona Spain, October 2006.

Sanderson C., Guenter. S. (2006). Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP), Sydney, Australia, 2006, pp. 482–491. Published by Association for Computational Linguistics (ACL).

Sayoud. H. (2010). Investigation of Author Discrimination between two Holy Islamic Books. IET (*ex-IEE*) Teknologia Journal. Vol. 1, Issue. 1, pp. X-XII, July 2010.

Tambouratzis G., Markantonatou S., Hairetakis N., Vassiliou M., Tambouratzis D. & Carayannis G. (2000). Discriminating the Registers and Styles in the Modern Greek Language . *Proceedings of the Workshop on Comparing Corpora* (held in conjunction with the 38th ACL Meeting), Hong Kong, China, 7 October. pp. 35-42.

Tambouratzis, G., Markantonatou S., Vassiliou M., Tambouratzis D. (2003). Employing Statistical Methods for Obtaining Discriminant Style Markers within a Specific Register**.** In Proceedings of the Workshop on Text Processing for Modern Greek: From Symbolic to Statistical Approaches (held in conjuction with the 6th International Conference of Greek Linguistics), Rethymno, Greece, 20 September 2003, pp. 1-10.

Wiki1. (2012). Quran. The free encyclopedia. Wikipedia, Last modified 2011, http://en.wikipedia.org/wiki/Quran

Wiki2. (2012). Hadith. The free encyclopedia. Wikipedia, Last modified 2011, http://en.wikipedia.org/wiki/Hadith

A framework for detecting Holy Quran inside Arabic and Persian texts

Mohsen Shahmohammadi[1], Toktam Alizadeh[2], Mohammad Habibzadeh Bijani[3],Behrouz Minaei[4]
[1] Islamic Azad University Tehran (North), Tehran, Iran,
[1,2,3,4] Computer Research Center of Islamic Sciences (Noor), Qom, Iran
[4] University of Science and Technology, Tehran, Iran
E-mails: mshahmohammadi@noornet.net, t_alizadeh20@yahoo.com, mhabibzadeh@noornet.net, b_minaei@iust.ac.ir

Abstract

This paper presents how to design and implement the Quranic intelligent engine to detect Quranic verses in the texts automatically. Process area of this system is in the scope of text mining processes and its operations are beyond the usual multiple patterns matching for reasons are explained in the paper. A new algorithm based on indexing text and patterns is designed in implementation of this system in which the main idea is to map text and patterns to some numerical arrays and process on them rather than the text. This algorithm detects Quranic verses in two stages. In the first stage, using a new index based exact multiple patterns matching algorithm, Quranic words in the text are detected and are converted to numerical arrays. In the second stage a filter is designed that by search on the arrays is able to detect the indexing sequence between arrays and determine whether these words are a part of Quran. Experimental results show that processing on numerical values rather than the text has a significant impact on increasing the performance of algorithm to be faster and more precise for detecting holy Quran phrases inside the texts.

Keywords: text mining, exact multiple pattern matching, intelligent detection of Quranic verse.

1. Introduction

Highlight Quranic phrases in written texts -by change the font of writing or using editing marks- is a subject that has long been of interest to researchers, authors and polygraphs. Because of researchers can extract different data from books by catalog these phrases and their statistical processing. Then they follow their research in selected areas according to their needs.

Perhaps at first glance, searching and marking Quranic text by a machine is seem a simple matter, but the difficulties can be received only after recognition of differences. For example, modify any of the following different cases or withdraw each of them can greatly influence the statistical results derived from the notation:

- A text has many similarities with a verse of Quran but has some differences in some ways like vowels and pronoun. This may not be a part of Quran!
- The text is quite similar to Quran, but in terms of its position in the text has no value to be introduced as a verse of Quran.
  There are some three-word phrases that are common in Arabic and there are also in Quran. But if these words appear in the text alone, there would be no value in terms of semantic to be detected as a verse of Quran. For example, “و هو علی”.
- Acceptable speed, precision and comprehensiveness for searching Quran in the text
  A system that is designed to detect Quran verses must have acceptable speed so it can replace humans.
  Along with speed, the system should have high precision in determining whether a text is part of Quran. The accuracy parameters have been fully considered in this system. Also about the comprehensiveness should be noted that this system isn’t considered any limiting hypothesis. The text can include Quranic phrases or not, may have or don’t have editing marking and words may have one or more various spelling. The text size can be small or very large.
- Possibility of existing various spelling of Quran phrases in a text.
  There are different spellings for writing Quran. It may be used different spellings for writing Quran verses in a book or one author uses their own spellings to write Quranic phrases which is different from conventional spellings. To solve this problem, a solution is that all spellings of the book convert to a single spelling and then search that text. But this issue can be problematic when the purpose of the book compilation or editing of a special heading of its, is to examine the difference of spellings (for example, ”almoqna fi rasme masahefe Alamsar” book from one author of fifth century). So any changes in text would be inconsistent with the author purpose and must be thought to another choice.
- Highlight “بسم الله” and praise and similar cases that are used at the beginning and end of books, letters and some texts as a Quranic text can be questioned.
- Some texts have no similarity with Quran and even are written with a non-Arabic language, but have a clear meaning associated with parts of Quran. However it may be established a mutual relationship between these texts and Quran.

The reset of this paper is organized as follows: Section 2 includes the surveys of related works and existing string matching. The next section contains details about the exact multiple pattern matching algorithm specifically for

Quran. We describe our idea and our proposed algorithm in section 4 and evaluate the performance of them in section 5. Finally, we draw conclusion in Section 6.

2. Background and related work

The main algorithm of the intelligent engine to detecting Quran verses is placed in the area of multiple strings matching algorithms in which the main problem is searching multiple strings in a text at the same time Bhukya(2011). Multiple pattern matching is the computationally intensive kernel of many applications including information retrieval and intrusion detection systems and … (Bhukya, 2011; Kouzinopoulos, 2011). So far, several algorithms have been applied to this problem, but they all have some differences with this algorithm and existing conditions on this which is expressed in continue. In addition, So far such this system dosn't have been implemented specifically for Quran in the world. Pattern matching algorithms can be divided into two general groups (Salmela,2007; Moraru,2011; Scarpazza, 2008):

- Tree-Based Algorithms

  These algorithms are included Aho-Corasick Aho (1975), Commentz-Walter Kouzinopoulos(2011), SBOM Allauzen(2001); (Jianlong,2011) and etc. the main solution of these algorithms is to build a tree of the patterns and search the text with the aid of the tree. The main problem with these algorithms is that the tree grows quite rapidly as the pattern set grows and require more storage space. So utilization of these methods is not suitable for large pattern like Quran that has 84000 words.

- Multiple Pattern Matching Based on Rabin-Karp

  The main idea of these algorithms is that given a text position and a filter can specify if there can be a match pattern at this position or not. In (Salmela, 2007) the hashes values in Rabin-Karp algorithm acts as a filter and Moraru (2011) uses a filter called feed-forward Bloom filter for this purpose. A good filter is fast and produces few false positives. In addition, a verifier is needed to distinguish between false and true positives.

With a closer look, many existing pattern matching algorithms are reviewed and classified in two categories.

- Exact string matching algorithm
- Inexact/approximate string matching algorithms

Exact string matching algorithm trying to find one or all exact occurrences of a string in a sequence. Some exact string matching algorithms are Naïve Brute force algorithm, Boyer-Moore algorithm Bhukya(2011), KMP Algorithm Singla(2012), Horspool, Rabin-Karp Singla(2012), Zhu-Takaoka and Quick-Search Bhukya(2011).

So, the algorithm of our system is placed in the area of the exact multi-pattern matching. These algorithms generally are classified in two categories (Kandhan, 2010): Shift table-based algorithms and automaton based algorithms.

3. Multi-Pattern Matching about Holy Quran

In intelligent engine to detecting Quran verses, the patterns are Quran words that are stored in a data base as a single word. Although the exact multi-pattern matching is formed the main core of this intelligent engine algorithm, but this engine in many aspects is different from other search engines that do multi-pattern matching. Among these differences can be pointed to the following cases:

- Search sequential patterns

  One of the unique issues in this system, that makes it more complicated than other multi-pattern matching algorithms is that the sequence of patterns is also a condition for the pattern searching. That is if a word of text matches with a pattern, the next word in the text must be matched with the next pattern in the pattern data base in order to detect fragments of verses in the text. So the problem changes from search pattern to search for sequential patterns, but this procedure doesn't matter in other multi-pattern matching algorithms.

- Different spellings for a pattern

  One word in Quran may have different spellings. That means that a pattern may appear with different forms which all of them should be recognized as a pattern.

- Very large volume data sets patterns in this algorithm compared other algorithm

  The Qur'an contains more than 88400 words which each of them makes up a pattern. In addition, this search engine will support all existing Quranic spelling and this increases the size of patterns about 3 times. Obviously, the algorithm used for multiple patterns matching in this system is very different from the algorithms used in small collection. However, the pattern matching algorithms on data sets with more than 10,000 patterns are identified as very large data sets (Salmela, 2007). Also usually data sets that are considered as very large data sets support up to 100,000 patterns. However, data sets (patterns) in this search engine are several times this number.

- Implementation exact multiple pattern matching system specifically for Quran

  As mentioned, although many algorithms have been implemented that do exact multiple pattern matching, but so far no system has been implemented that does this matching specifically for Quran as in this system have been considered. Although various search engines have been implemented for Quran that search a word, a verse or a part of a verse accurately and address them in the text but neither does operations of this system to be able to detect Quranic phrases contained in the text automatically. For example, in (mohammadi nasiri, 2008) a Quranic search engine is implemented for web pages. Operation of this system is such usually searches that gets a verse and submits a report on presence or absence of the verse in that web page. Also in (Hammo, 2007) a Quranic search system is implemented that searches Quranic verses without dependence to vowels and diacriticals. However, in both these systems and other Quranic search systems (with slight differences in functionality) verse or part of the verse is given as input to the system and the search is done, while in our system no verse is given as input to the system, but all the verses in the text are detected automatically.

Due to expressed content and considering the advantages and disadvantages of existing methods, we design a new algorithm based on indexing text and patterns that detects Quranic verses in text in two stages. Our algorithm is explained in continue.

4. Proposed algorithm

The algorithm of intelligent engine to detect Quran verses in texts is consists of four steps which are described respectively:

1- making Quranic data base.
2- Initial processing of the input text.
3- Identify Quranic Words.
4- Search for sequential patterns and detection of verses.

As is common in string matching algorithms, in our algorithm two primary steps can also be considered as preprocessing and two next steps as main search operations and multiple string matching.

4.1. Making Quranic data base

This is the first step. In this step, full text of Quran is stored in an array Verbatim (word by word) and consecutive in order to belong a unique index to each word of Quran. The purpose of this unique indexing with this knowledge that many of the words of Quran are duplicate, is determined in continue.

4.2. Initial processing of the input text

At this step, first additional characters such as editing marks and vowels and etc are removed from the text. Then an index is assigned for each of remaining words in the input text, in order to their positions in the text be determined.

4.3. Identify Quranic Words

In this step each word of text is compared with patterns in data base patterns respectively and (if any matching is existed) the index of matching word is stored in an array that is allocated to its. (May be more than one pattern is matched with desired word. Because of for example “ﷲ” exists in different parts of Quran, but its index at each location is different from other locations). So every word of text is mapped to an array of numbers. It should be noted that it is possible that the word doesn't match with any of the patterns but it matches with a spelling of a pattern. For this reason in the implementation of this search engine is used a method that all words that have different spellings are also found and indexed at this step. This procedure will continue until don't find any matching pattern for input word in pattern database. Thus we have several arrays with the same number of previous words that each of them specifies the locations of that word in Quran.
For example, suppose that the first words of input text is included: “اللهُ الذّی هو الحیّ القیّوم لا تأخذهُ سنة و لا نَوم ....”. Input words are processed respectively until we get the non-Quranic word. Here it is assumed that the word after “نوم” word is a non-Quranic word. So the words are processed from the beginning to the end of the text and following arrays will be make for them.

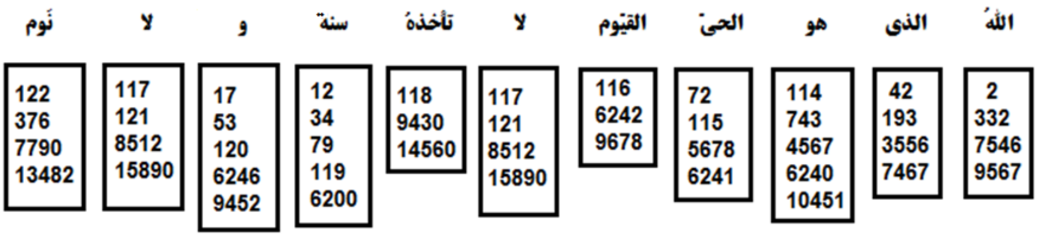


Figure 1: numerical arrays of words indexes in Quran

4.4. Search for sequential patterns and detection of verses

After the previous steps we now have a text that all of its words are Quranic words but may not necessarily a verse of Quran. To detect this, we search a sequence indexing between these arrays. At this step, the longest found matching string is returned as a verse or part of Quran. In the above example two index sequences are specified with the array processing.

1- «هو الحی القیوم»: part of verse 3 of Ale Emran sura.

2- «هو الحی القیّوم لا تأخذهٌ سنة و لا نَوم»: part of verse 255 of Baqarah sura (no 2).

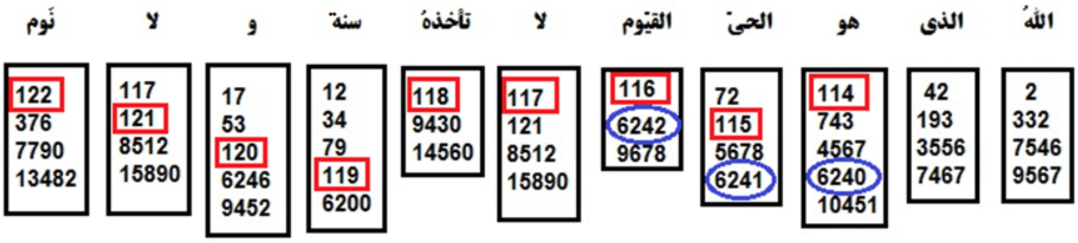


Figure 2: sequence indexing between some of arrays and detecting verse of Quran

In the last part of this step, Quranic verses are retrieved using patterns index (That was performed in pre-processing step). Then by the words index in the input string, the place of Quranic verses in the text is also determined and these words are highlighted in some ways (For example, text color change, parenthesis, ...) to indicate that the script is quoted from Quran.

The advantage of our method is that since we sort the index arrays of the words, we can use easily optimized search algorithms like binary search.

5. Implementation and results

We have done different tests to evaluate the performance of our proposed algorithm. At the first, test data is described. Then evaluation criteria are introduced. At last experimental results is presented and analyzed. The presented results are obtained from the average of different result tests.

5.1. Data sets

In this paper two types of resources are used to evaluate the results of proposed algorithm.

1- The resources that have high frequency of Quranic verses.
2- The resources in which there are fewer Quranic verses

The reason of these classifications is due to the purposes of searching Quranic verses by these two type categories are different. For example, in the texts that have low frequency of Quranic verses, the basic problem is not to find Quranic verses, but the goal is to identify large sections of text that there is no verse in them. Thus it will avoid the additional search in pages which have no Quranic verse.

5.2. Evaluation criteria

The most commonly used criteria to evaluate systems performance include: Recall, Precision and F-Score. In our algorithm recall for a verse is the number of words that are correctly detected as Quran compared to the total number of Quran words in the text. Precision for a verse is the number of words that are correctly detected as Quran compared to the total number of words that have been identified as Quran. F-Score is the harmonic mean of precision and recall and shows the impact of the precision and recall simultaneously.

5.3. Experimental results

Two types of resources have been used for evaluating this system. One, the resources with high frequency of Quranic verses and other the resources with low frequency of Quranic verses. Experimental results are presented separately for each of them. It should be noted that there will be two modes for each word of a text that is checked by this system: Each word is detected as a Quranic or non-Quranic word. Accordingly, two criteria can be defined for evaluating the system.

1- The words that are detected as Quran
2- The words that are not detected as Quran

In this paper both cases is considered and the geometric mean of these two is also calculated for the overall conclusion. The tables 1 and 2 show the results of these experiments.

Words Type	Precision	Recall	F- Score
The words are detected as Quran	0.866788	1	0.928641
The words are not detected as Quran	0.96242	1	0.98085
Geometric mean	0.92864	1	0.95403

Table 1: The source contains 2418 words and high frequency of Quranic verses

Words Type	Precision	Recall	F- Score
The words are detected as Quran	0.95383	0.95833	0.88010
The words are not detected as Quran	0.99555	0.97704	0.98261
Geometric mean	0.89547	0.96760	0.93014

Table 2: The source contains 3802 words and low frequency of Quranic verses

6. Conclusion

In this paper a new intelligent algorithm to detect Holy Quran in digital texts is presented. The proposed algorithm is a kind of index-based exactly multiple string matching algorithms. But some special conditions of this problem such that searching for sequential patterns makes it more complicated than the usual exact multiple pattern matching. Experimental results show that the main idea of the algorithm of mapping patterns to numerical arrays and process on those number arrays is a significant impact on increasing the performance of algorithm to be faster and more precise. So that all Quranic verses in the text are detected by the algorithm with high precision.

7. Acknowledgments

This paper is developed by supports of the Computer Research Center of Islamic Sciences (Noor).

8. References

A. Aho, M. Corasick; Bell Laboratories. "Efficient String Matching:An Aid to Bibliographic Search", in *Communications of the ACM*, vol. 8, 1975, pp. 333-340.

C. Allauzen, M. Crochemore ; M. Raffinot, "Efficient Experimental String Matching by Weak Factor Recognition", in *Proc. 12th Annu. Symp. on Combinatorial Pattern Matching*, Jerusalem, July 1–4, 2001, pp. 51-72.

Ch S. Kouzinopoulos, P D. Michailidi, K G. Margaritis, "Parallel Processing of Multiple Pattern Matching Algorithms for Biological equences: Methods and Performance Results*", Systems and Computational Biology - Bioinformatics and Computational Modeling,* ISBN 978-953-307-875-5, pp. 161-182, InTech, September 2011.

Hammo, B., Sleit, A; El-Haj, M.. *"Effectiveness of Query Expansion in searching the Holy Quran". In Proceeding of the Second International Conference on Arabic Language Processing CITALA'07,* pages 1-10, Rabat, Morocco 2007. CITALA.

I Moraru, David G. Andersen*, "Exact Pattern Matching with Feed-Forward Bloom Filters". 2011 by SIAM.*

L Salmela,*"Multi-Pattern String Matching with Very Large Pattern Sets". ACM Journal of Experimental Algorithmic, Volume 11, 2006, November 1st 2007.*

Mojtaba mohammadi nasir, "Some issues in the design of Quranic web search engine", *Master's thesis presented in computer software engineering, Under doctor Mohammad Ghodsi, Sharif university of technology, Department of computer engineering, 2008.*

N Singla, D Garg, "String Matching Algorithms and their Applicability in various Applications", *International Journal of Soft Computing and Engineering (IJSCE),* ISSN: 2231-2307, Volume-I, Issue-6, pp. 218-222, January 2012.

Raju Bhukya, DVLN Somayajulu, "Index Based Multiple Pattern Matching algorithm using DNA Sequence and pattern count", *International Journal of Information Technology and Knowledge Management*, Volume 4, No. 2, pp. 431-441, July-December 2011.

Raju Bhukya, DVLN Somayajulu, **"***Exact Multiple Pattern Matching Algorithm using DNA Sequence and Pattern Pair". International Journal of Computer Applications (0975 – 8887).* Volume 17– No.8, March 2011.

R Kandhan, N Teletia, J M. Patel, "*SigMatch: Fast and Scalable Multi-Pattern Matching*", *Proceedings of the 36th International Conference on Very Large Data Bases,* Vol. 3, No. 1, September, 2010, Singapore.

Scarpazza, D.P., Villa, O.; Petrini, F., *"High-speed string searching against large dictionaries on the Cell/B.E. Processor ". Parallel and Distributed Processing, 2008. IEEE International Symposium on.* ISSN: 1530-2075, pp. 1 – 12. 14-18 April 2008 , Miami, FL.

T Jianlong, L Xia; L Yanbing; L Ping "Speeding up Pattern Matching by Optimal Partial String Extraction", in *The First International Workshop on Security in Computers, Networking and Communications (INFOCOM WKSHPS), 2011 IEEE* , E-ISBN: 978-1-4577-0248-8, Shanghai , 23 June 2011 , pp. 1030 – 1035.

Letter-to-Sound Rules for Gurmukhi Panjabi (Pa): First step towards Text-to-Speech for Gurmukhi

Gurpreet Singh
Centre for Language and Communication Studies
SLSCS, Trinity College Dublin, Ireland
gursainipreet@gmail.com

Abstract

This article presents the on-going work to develop Letter-to-Sound rules for Guru Granth Sahib, the religious scripture of Sikh religion. The corpus forming the basis for development of the rules is taken from EMILLE corpora. Guru Granth Sahib is collection of hymns by founders of Sikh religion. After presenting an overview of Guru Granth Sahib and IPA representation in section 1 and Text-to-Speech in section 2, Letter-to-Sound rules developed will be presented in section 3. This paper will close with final discussion and future directions in section 4. The work presented stand at the development stage and no testing or experiment have so far been performed. The intention is to develop the Text-to-Speech for Punjabi language after developing it for limited set of language available in Guru Granth Sahib.

Keywords: Rule Based Letter-to-Sound System, Gurmukhi Punjabi, Sikhism

1. Guru Granth Sahib, EMILLE and Panjabi (Pa)

1.1 Sri Guru Granth Sahib (SGGS): It is a voluminous text of 1430 pages (called angs), compiled and composed during the period of Sikh Gurus, from 1469 to 1708. It is a collection of hymns (shabad or "Baani"). It is written in the Gurmukhi script, predominantly in archaic Punjabi along with languages including Braj, Punjabi, Khariboli (Hindi), Sanskrit, regional dialects, and Persian.

1.2 EMILLE: The EMILLE corpora consist of fourteen languages, namely Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu, and Urdu. The monolingual written Punjabi corpus consists of around 15,600,000 words (Baker et al., 2004). Apart from the monolingual corpora for all 14 languages, there are spoken and parallel corpora for some of the 14 languages. Work presented in (Baker et al., 2004; McEnery et al., 2004) can be consulted for further information on EMILLE corpora. The part of Punjabi Corpus being used for the work being presented is in writtengurmukhi/shreegurugranthsahib and the corpus file is named pun-w-relig-shreegurugranthsahib.txt. The text in the file two-byte Unicode encoded text.
A copy of EMILLE can be obtained from ELRA/ELDA (W0037, 2010) free of charges for non-commercial research purposes.

1.3 Gurmukhi Punjabi (or Panjabi) (Pa)
Punjabi (or Panjabi) is a language written in two different scripts namely Gurmukhi translated literally as "from the mouth of the gurus" and Shahmukhi. It is spoken in the states of Punjab in both India and Pakistan. For Sikhs, the Punjabi (Gurmukhi) is the official language for all ceremonies and rituals. Punjabi (Shahmukhi) is the most spoken language in Pakistan. Gurmukhi Punjabi is also known as Eastern Punjabi. It has around 28,163,000 (Lewis, 2009) native speaker belonging mainly to Punjab (India).
Punjabi is an Indo-Aryan language belonging to the Indo-European language family's Indo-Iranian subgroup. It has been classified as belonging to the CENTRAL group under INNER sub-branch of Indo-Aryan languages under NIA sub classification scheme (Masica, 1991; p. 449). Nagari or Devanagari script is also used to write Punjabi in the areas of Jammu (Cardona & Jain, 2003; p. 53).

1.3.1. Gurmukhi Script
Gurmukhi script is an abugida writing system, where each consonant has an inherent vowel (/ə/) modifiable using vowel symbols which can be attached to the relevant vowel-bearing consonant. Gurmukhi Punjabi has 35 native characters and another 6 characters to accommodate sounds from foreign languages. These characters represent 3 vowel characters, 2 fricatives, 25 consonants, 5 semi-vowels. In addition there are nine vowel symbols, two symbols for nasal sounds and one symbol which duplicates the sound of any consonant.
The Punjabi phoneme inventory has twenty-five consonant phonemes, ten vowel phonemes, three tones (High, Mid, Low), and seven diphthongs. A number of non-native speech sounds are also found, however, these sounds occur in loan words only (mostly Persian and Arabic loans). Punjabi is the only tone language in the Indo-European language family, making it of considerable interest to both phonologists and historical linguists (UCLA, 2010).
Gurmukhi script has quite a different structure and system as compared to all the other Indian Scripts. This has been attributed to two main facts, as suggested in (Cardona& Jain, 2003; p. 83):

1. The tonal system and some other phonetic features,
2. Different cultural and historical circumstances.

All of the Sri Guru Granth Sahib Ji's 1430 pages are written using Gurmukhi script. It has a total of 398,697 words with a total of 29,445 unique dictionary words. Many of these words have been used only once.

1.4 IPA representation

International Phonetics Association (IPA) letters are used in the system to represent the sounds. Appendix B shows the Gurmukhi letters and symbols and their corresponding IPA letters.

Before actually moving on to the system and rules developed in section 4, sections 2 and 3 will explain a little on Text-to-Speech and Letter-to-Sound.

2. Text-to-Speech

A computer based program that is capable of reading any text is known as a Text-to-Speech synthesizer. Unlike Voice Response Systems, which are applicable to limited vocabulary and very restricted sentence structures, for TTS systems it is impossible to know all the vocabulary in advance. Taking this into account Dutioit (1997) has defined Text-to-Speech as:

> "automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter"

Possible uses of such TTS systems include, but are not limited to:

1. Telephone and communication services
2. Language education
3. Aid for handicapped persons
4. Talking books and toys
5. Multimedia, human computer interaction.

A complete Text-to-Speech system includes more than just Letter-to-Sound or Grapheme-to-Phoneme components. The main components of a TTS system at very general level include:

1. Natural Language Processing (NLP)
2. Digital Signal Processing (DSP)

NLP component of the TTS produces the phonetic transcription and prosody of the text to be read, which is fed to DSP component to be converted into speech. In this work the topic of concern is generation of phonetic transcription of the text to be read. The text of concern is not the complete domain of Gurmukhi Punjabi, but it is limited to the text in Guru Granth Sahib Ji (also called 'Gurbaani') as explained in the section 1.

3. Letter-to-Sound Rules

In the Letter-to-Sound (LTS) module, automatic phonetic transcription of the text to be spoken is done. Broadly speaking the LTS module can be classified into two categories:

1. Dictionary-based: It consists of storing maximum phonological knowledge about the language in the lexicon to be looked up for LTS component to work.
2. Rule-based: It uses the rules for generating the phonetic transcription from the text. In this work, a rule based LTS system will be developed. This is a language dependent approach and can only be used for Gurmukhi Punjabi for Guru Granth Sahib Ji. In future, this system can be easily adapted for general Punjabi language.

Letter-To-Sound rules are always required, even if there is very good dictionary or lexicon available. More about the comparability between the two is given in section 1.3. The system presented is a pure rule-based system, developed using Java.The mapping is from Gurmukhi to International Phonetic Association (IPA) alphabets as explained in section 1.4 and general mapping can be seen in Appendix B. The complete system and the rules will be explained in the next section.

4. The System

Gurmukhi Panjabi can be classified as phonologically transparent or shallow orthography (Raghallaigh, 2010; p.64) writing system where the symbol (letter)-sound correspondences are very regular and has mostly one-to-one letter-to-sound mapping, unlike English. As Hand-written rules work well for the languages with transparent orthography, the system presented uses hand-written rules.

Most of the spoken features of written text can be captured by a very simple rule base. The rule based system for converting letters to the sounds will be presented, starting with a very simple system. One thing that needs to be mentioned before actually starting with the description of the system and the rules is that the system being presented is for the language present in the text of Guru Granth Sahib Ji and is not the modern Punjabi. The difference between the two will be explained in section 4.1 and system and rules in section 4.2.

4.1. Difference between the Language of Guru Granth Sahib Ji and Modern Punjabi

As stated in section 1, Shiri Guru Granth Sahib Ji is a voluminous text comprising of 1430 pages. The text is a collection of hymns written by six Sikh gurus and other great saints, or Bhagats, including those of the Hindu and Muslim faith. Although written in the Gurmukhi script, the text are in different languages including Braj, Punjabi, Khariboli (Hindi), Sanskrit, Arabic, Sindhi, Lehndi, Dakhni, Bengali, Marathi and Persian, given the generic title of Sant (Saint) Bhasha (language) (meaning the language of the saints). The text contains the hymns written by a total of 42 writers. The contribution of each of the writers varies a great deal in the amount of text included.

The variety in the languages and authors has contributed towards the unique language used in Gurbaani. Another reason for the uniqueness of the language is due to the fact that the original scripture was written as a continuous text without any spaces between the words. The differences between the two, Modern Gurmukhi and Gurmukhi in Gurbaani, are as follows:

1. The Gurmukhi script used in Shiri Guru Granth Sahib Ji uses the only first 35 characters of the 41 characters used in modern script.
2. There is no use of Adhdhak (/ੱ/) in Shiri Guru Granth Sahib Ji. In modern writing this symbol is used frequently for the purpose of doubling the sound.
3. No Punctuation marks were used; instead vowel symbols fulfilled the purpose.
4. The use of vowel sounds (ਿ) and (ੁ) at the end of the word, with some exceptions, does not form the syllable. These were used for grammatical reasons.
5. Use of characters (ਚ, ਟ, ਤ, ਨ, ਯ, ਵ) in the foot of other letter and symbols and half characters is not seen in the modern writing.
6. Bindi (ਂ) : It has been written on the left hand side of vowel signs, which is not permissible in modern language.

4.2. Rules and the system

This section will explain the rules used in the development of the Letter-To-Sound system. These rules are inspired by the work done by Singh (2007). Apart from Gurmukhi being a phonologically transparent writing system, the other reason to use hand-written rules was that the task in hand is a closed domain task. All the possible words being available, any and every words' pronunciation can be captured using the rules.

In the simplest rule based system developed to start with, Gurmukhi Panjabi being a phonologically transparent writing system, corresponding arrays of Gurmukhi alphabets and vowels symbols were generated and the only rule was to replace the Gurmukhi symbol by the Corresponding IPA symbol. This system, as expected, had several shortcomings as a LTS system for the Gurbaani. When results were compared to the words in the corpus, several repeated errors or shortcomings were noticed.

Further rules were developed and added to the final system to deal with these. These rules are as explained below:

1. **ੴ**: '**ੴ**' is a special symbol which is always pronounced as /ɪkoəŋ̍kɑr/. This being the first and very frequent character in the corpus forms the first rule in the rule base. In fact, it has been repeated 568 times in the corpus.

2. Numbers: In the corpus, when the numbers are used for information purpose, they need not be pronounced at all. Although numbers as high as 243 can be found in the corpus, but based on the rules for whether or not to pronounce a number. The highest number that can be pronounced is found to be 16. The numbers can be pronounced in both their counting form as well as the adjective form. After consulting the corpus being used it was found that the decision about whether to pronounce the number or not and which form to pronounce depends on the word preceding the number. The rules for the pronunciation of numbers can be seen in the appendix A along with other rules. Zero as a single digit number is not found in the corpus so there is no rule for zero.

3. Single Consonants: There were words composed of single consonant alphabet found regularly in the corpus. 7119 such words were found in the corpus comprising of only four consonants. These consonants were 'ਤ ', 'ਨ ', 'ਕ' and 'ਚ'. The rule to pronounce these words is to add /ɑ/ sound at the end of the single consonant.

4. Consonants of fourth column (refer to alphabet arrangement in Appendix A): The only characters to have more than one sound were found to be the characters 'ਘ', 'ਝ', 'ਢ', 'ਧ' and 'ਭ'. All of these belong to the fourth column for five rows, second to sixth, of Gurmukhi alphabets. These characters, when found at the starting position of the word, are pronounced using their own sound. The rule for the additional sound is that when they are found at any other position in the word, then they are pronounced as the sound of third alphabet of their corresponding row

5. (ਿ)('ɪ') and (ੁ)('ʊ') at the end of the word: As mentioned in section 4.1, these two vowel sounds serve more purposes than merely being the vowel sounds in the Gurbaani text. They serve grammatical purposes as case, gender and end of word markers. Depending on the accompanying alphabet, these have to be dealt with differently. Sometimes the sound is not pronounced at all and when pronounced it has to be changed or can be the exact pronunciation. All the rules to handle these sounds are as below:

(a) Single consonant: The sound of 'ɪ' changes to 'e' and 'ʊ' to 'o', when these vowel sounds are found with a word having only one consonant.

(b) With consonant in longer words: The sounds 'ɪ' and 'ʊ' are not generally pronounced if these are at the end of the word. The exceptions to this case are when the sounds are found with 'ਹ' ('ɦ') and 'ਯ' ('j'). With 'j' the sounds are always pronounced while with 'ɦ', the sound of 'ɪ' changes to 'e' and that of 'ʊ' to 'o'. Even when 'ਿ' is found as 'ਇ', the same change applies and 'ਇ'and is pronounced as 'e' instead of 'ɪ''.Similar is the case with 'ਉ' which is pronounced as 'o' instead of 'ʊ'.

6. 'ɪ' and 'ʊ' followed by vowel 'ਅ' ('/ə/') as the word ending: The sounds 'ɪ' and 'ʊ' followed by 'ਅ' ('@') is changed to 'e' and 'o' respectively and the last vowel sound /ə/ is dropped.

7. Nasal sounds: In Gurmukhi, the two symbols ' ਂ ' and 'ੰ ' are used to nasalize the sound of the symbol with which they are associated. The second of the symbols 'ੰ ', serves the purpose of duplication of the sound as well. ' ਂ ' will nasalize the sound of the consonant or vowel after which it is written. The sound rules for 'ੰ ' are as follows:

(a) When used with single consonant it produces soft 'n' sound.

(b) When found at the end of the word it produces the sound of 'ŋ̍'.

(c) When coming before the nasal sounds, the letters in the last column from row two to six, it duplicates the sound of the letter with which it occurs.

(d) In all other cases it will produce a nasal sound depending on the row to which the associated letter belongs. The nasal sound is from the same row as is the associated letter for the rows having the nasal sounds. For the remaining two rows, first and seventh, the nasal sound

‘n’ will be pronounced.
8. ‘॥’: This is another special character used in the corpus. It is used as end of the line marker, so it is never pronounced.

All the rules are described in the appendix A.

5. Discussion

This work has developed a rule-based Letter-to-Sound system for Gurmukhi used in Shiri Guru Granth Sahib ji. This is the first step towards developing the complete Text-to- speech system for the same language. There is a room for improvements and further enhancements. This work so far has been focused on the rules which can produce a system capable of handling the text in the corpus. The development phase for each rule was also a test bed as it was ensured that the new rules integrated with the existing ones.

Immediate future work will be the analysis of rules and testing of the system. The other parts of the used corpora are intended to be used for further developments. The areas not addressed by this work include reduction of inherited vowel sound /ə/ and lexical stress to name a few. For a complete TTS system there are further resources required which include tools for morphological analysis and part-of-speech tagging, which are not available at the time. In the end it is hoped that the work can inspire the others to develop this work towards a successful TTS system for Gurmukhi Punjabi.

6. Acknowledgements

Sincere thanks to Dr. Elaine Ui Dhonchadha for her support and advice during this work.

7. References

Panjabi, eastern@ONLINE, 25 - Nov 2010. http://www.ethnologue.com/show_language .asp?code=pan.

The EMILLE Project, 21 - Dec 2010. http://www.emille.lancs.ac.uk/.

UCLA: Language Materials Project, 21 - Dec 2010. http://www.lmp.ucla.edu/Profile.aspx? LangID=95&menu=004.

W0037: The EMILLE/CIIL Corpus, 21 - Dec 2010. http://www.elda.org/catalogue/en/ text/W0037.html.

Baker,P.; Hardie, A.; McEnery, T.; Xiao R.;Bontcheva, K.; Cunningham,H.; Gaizauskas,R.; Hamza,O.; Maynard, D.; Tablan, V.; Ursa,C.; Jayaram, B. D. and Leisher. M. (2004). Corpus linguistics and south asian languages: Corpus creation and tool development, In *Literary and Linguistic Computing, volume 19,* pp. 509 – 524.

G. Cardona and D. Jain (Eds). (2003). The Indo-Aryan, London: Languages. Routledge.

Dutoit, T. (1997). High-quality text-to-speech synthesis: An overview. In *Electrical and electronics engineering, (Volume 1).* pp 25–36.

Lewis M. P. (Eds). (2009). *Ethnologue: Languages of the World (16th Edition),* Dallas, Tex: SIL International

Masica, C. P.; (1991). *The Indo-Aryan Languages.* Cambridge: Cambridge University Press.

McEnery, A.; Baker, P.; Gaizauskas, R. and Cunningham, H. (2004). Emille: Building a corpus of south asian languages. In *Literary and Linguistic Computing, (volume 19).* pp 509 – 524.

Raghallaigh, B. O. (2010). Multi-Dialect phonetisation for Irish text-to-speech synthesis: a modular approach. PhD thesis. School of Linguistics, Speech and Communication Sciences.

Singh, A. (2007). *Gurbaannee's Significance and PronunciationA Simplified Guide (second edition).* Sikh Youth Federation.

APPENDIX A

The complete rule set can be found in AppendixA.pdf at https://sites.google.com/site/gursainipreet/work-experience-and-projects

APPENDIX B

The Gurmukhi to IPA mapping can be found in AppendixB.pdf at https://sites.google.com/site/gursainipreet/work-experience-and-projects

Style of Religious Texts in 20th Century

Sanja Štajner, Ruslan Mitkov

Research Institute in Information and Language Processing
University of Wolverhampton, UK
S.Stajner@wlv.ac.uk, R.Mitkov@wlv.ac.uk

Abstract

In this study, we present the results of the investigation of diachronic stylistic changes in 20th century religious texts in two major English language varieties – British and American. We examined a total of 146 stylistic features, divided into three main feature sets: (average sentence length, Automated readability index, lexical density and lexical richness), part-of-speech frequencies and stop-words frequencies. All features were extracted from the raw text version of the corpora, using the state-of-the-art NLP tools and techniques. The results reported significant changes of various stylistic features belonging to all three aforementioned groups in the case of British English (1961–1991) and various features from the second and third group in the case of American English (1961–1992). The comparison of diachronic changes between British and American English pointed out very different trends of stylistic changes in these two language varieties. Finally, the applied machine learning classification algorithms indicated the stop-words frequencies as the most important stylistic features for diachronic classification of religious texts in British English and made no preferences between the second and third group of features in diachronic classification in American English.

Keywords: stylometry, language change, text classification

1. Introduction

According to Holmes (1994), style of the text could be defined "as a set of measurable patterns which may be unique to an author". Štajner and Mitkov (2011) further amended this definition to the scope of the language change by defining style "as a set of measurable patterns which may be unique in a particular period of time" and tried to examine whether "certain aspects of the writing style used in a specific text genre can be detected by using the appropriate methods and stylistic markers". In that study, they investigated only four stylistic features over the four main text genres represented in the 'Brown family' of corpora – Press, Prose, Learned and Fiction. In this study, we followed their main ideas and methodology but focused only on the genre of religious texts. We made a more in depth analysis of stylistic changes by using more features (a total of 146 features) and applied machine learning techniques to discover which features underwent the most drastic changes and thus might be the most relevant for a diachronic classification of this text genre.

1.1. Corpora

The goal of this study was to investigate diachronic stylistic changes of 20th century religious texts in British and American English and then compare the trends of reported changes between these two language varieties. Therefore, we used the relevant part (genre D – Religion in Table 1) of the only publicly available corpora which fulfills both conditions of being diachronic and comparable in these two English language varieties – the 'Brown family' of corpora. The British part of the corpora consists of the following three corpora:

- The Lancaster1931 Corpus (BLOB),
- The Lancaster-Oslo/Bergen Corpus (LOB)
- The Freiburg-LOB Corpus of British English (FLOB).

These corpora contain texts published in 1931±3, 1961 and 1991, respectively. The American part of the 'Brown family' of corpora consists of two corpora:

- The Brown University corpus of written American English (Brown)
- The Freiburg - Brown Corpus of American English (Frown).

These two corpora contain texts published in 1961 and 1992, respectively. Four of these corpora (LOB, FLOB, Brown and Frown) are publicly available as a part of the ICAME corpus collection[1] and they have been widely used across the linguistic community for various diachronic and synchronic studies as they are all mutually comparable (Leech and Smith, 2005). The fifth corpus (BLOB) is still not publicly available, although it has already been used in some diachronic studies, e.g. (Leech and Smith, 2009).

As the initial purpose of compiling the Brown corpus was to have a representative sample of 'standard' English language (Francis, 1965), the corpus has covered 15 different text genres, which could further be clustered into four main text categories – Press, Prose, Learned and Fiction (Table 1). The other four corpora which were compiled later (LOB, FLOB, Frown and BLOB) shared the same design and sampling method with the Brown corpus, thus making all five of them mutually comparable.

The genre which is relevant for this study – genre D (Religion), belongs to the broader Prose text category (Table 1). To the best of our knowledge, this genre has never been used in any diachronic study on its own, instead it was always included as a part of the Prose category, together with

[1] http://icame.uib.no/newcd.htm

Category	Code	Genre
PRESS	A	Press: Reportage
	B	Press: Editorial
	C	Press: Review
PROSE	D	Religion
	E	Skills, Trades and Hobbies
	F	Popular Lore
	G	Belles Lettres, Biographies, Essays
	H	Miscellaneous
LEARNED	J	Science
FICTION	K	General Fiction
	L	Mystery and Detective Fiction
	M	Science Fiction
	N	Adventure and Western
	P	Romance and Love Story
	R	Humour

Table 1: Structure of the corpora

the other four Prose genres (E–F). Although this genre contains only 17 texts of approximately 2,000 words each, the texts were chosen in the way that they cover different styles and authors of religious texts. For instance, in the Brown corpus, 7 of those texts were extracted from books, 6 from periodicals and 4 from tracts[2]. The full list of used texts and the authors for each of the four corpora could be found following the links given in their manuals[3]. Although the size of the corpora used in this study (approx. 34,000 words in each corpus) is small by the present standards of corpus-based research, it is still the only existing diachronic comparable corpora of religious texts. Therefore, we find the results presented in this paper relevant though we suggest that they should be considered only as preliminary results until a bigger comparable corpora of religious texts become available.

1.2. Features

In this study, we focused on genre D (Religion) of the four publicly available parts of the 'Brown family' of corpora and investigated diachronic stylistic changes in British (using the LOB and FLOB corpora) and American (using the Brown and Frown corpora) English. As the LOB and FLOB corpora cover the time span from 1961 to 1991, and the Brown and Frown corpora from 1961 to 1992, we were also able to compare these diachronic changes between the two language varieties in the same period 1961–1991/2. In both cases, we used three sets of stylistic features. The first set contains features previously used by Štajner and Mitkov (2011):

- Average sentence length (ASL)
- Automated Readability Index (ARI)
- Lexical density (LD)
- Lexical richness (LR)

Average sentence length has been used as a feature for stylistic categorisation and authorship identification since 1851 (Holmes, 1998; Gamon, 2004). It is calculated as the total number of words divided by the total number of sentences in the given text (eq.1).

$$ASL = \frac{total\_number\_of\_words}{total\_number\_of\_sentences} \qquad (1)$$

Automated Readability Index (Senter and Smith, 1967; Kincaid and Delionbach, 1973) is one of the many readability measures used to assess the complexity of the texts by giving the minimum US grade level necessary for its comprehension. McCallum and Peterson (1982) have listed it among eleven most commonly used readability formulas of that time, which was probably related to the fact that it is very easy to be computed automatically. Unlike the other readability indexes which usually require the number of syllables in text (difficult to compute automatically with a high precision), ARI only requires the number of characters (c), words (w) and sentences (s) in the given text (eq.2).

$$ARI = 4.71\frac{c}{w} + 0.5\frac{w}{s} - 21.43 \qquad (2)$$

Lexical density has already been in use as a stylistic marker in, e.g. (Ule, 1982) and for dating works in (Smith and Kelly, 2002). It is calculated as the ratio between the number of unique word types and the total number of tokens in the given text (eq.3). Therefore, a higher lexical density would indicate a wider range of used vocabulary.

$$LD = \frac{number\_of\_unique\_tokens}{total\_number\_of\_tokens} \qquad (3)$$

However, as lexical density counts morphological variants of the same word as different word types, Corpas Pastor et al. (2008) suggested that instead of lexical density, another measure – lexical richness, should be used as an indicative of the vocabulary variety. The lexical richness is computed as the ratio between the number of unique lemmas and the total number of tokens in the given text (eq.4).

$$LR = \frac{number\_of\_unique\_lemmas}{total\_number\_of\_tokens} \qquad (4)$$

This second measure does not take into account different morphological counts of the same word as different word types and therefore, Corpas Pastor et al. (2008) believed that it would be a more appropriate indicative of the vocabulary variety of an author.

The second set of features contains nine different part-of-speech frequencies:

- Nouns (N)
- Pronouns (PRON)
- Determiners (DET)
- Prepositions (PREP)
- Adjectives (A)
- Adverbs (ADV)
- Coordinating conjunctions (CC)

[2] http://icame.uib.no/brown/bcm.html

[3] http://khnt.aksis.uib.no/icame/manuals/index.htm

- Subordinating conjunctions (CS)
- Verbs (V)
- Present participles (ING)
- Past participles (EN)

The third set of features were the following 123 stop words (Table 2), based on the 'Default English stopwords list'[4]. In our case, as the used parser treats negative contractions as separate words, transforming for instance *couldn't* into two words: *could* and *not*, we excluded all the words with negative contractions from the original list.

a, about, above, after, again, against, all, am, an, and, any, are, as, at, be, because, been, before, being, below, between, both, but, by, could, did, do, does, doing, down, during, each, few, for, from, further, had, has, have, having, he, her, here, hers, herself, him, himself, his, how, i, if, in, into, is, it, its, itself, me, more, most, my, myself, no, nor, not, of, off, on, once, only, or, other, ought, our, ours, ourselves, out, over, own, same, she, should, so, some, such,than, that, the, their, theirs, them, themselves, then, there, these, they, this, those, through, to, too, under, until, up, very, was, we,were, what, when, where, which, while, who, whom, why, with,would, you, your, yours, yourself, yourselves.

Table 2: Stop words

2. Related Work

Since the 1990s, when the FLOB and Frown corpora were compiled, a great amount of diachronic studies in both American and British English has been conducted using the 'Brown family' of corpora (Brown, Frown, LOB and FLOB). Mair et al. (2003) investigated diachronic shifts in part-of-speech frequencies in British English, reporting an increase in the use of nouns and a decrease in the use of verbs in the Prose text category in the period 1961–1991. Štajner and Mitkov (2011) compared the diachronic changes in the period 1961–1991/2 between British and American language varieties, taking into account four stylistic features: average sentence length (ASL), Automated Readability Index (ARI), lexical density (LD) and lexical richness (LR). Their results indicated increased text complexity (ARI) in the Prose genres of British English, and increased lexical density and lexical richness in the Prose genres of both language varieties over the observed period (1961–1991/2).

It is important to emphasise that in all of these previous diachronic studies conducted on the 'Brown family' of corpora, the authors did not differentiate across different genres in the Prose category (among which is the relevant genre D – Religion), but they rather examined the whole Prose text category together. As the Prose category is comprised of five rather different text genres (Table 1 in Section 1.1), we cannot know whether their findings would stand for the Religious texts (genre D) on its own. Therefore, we included all of these features in our study. This way, by comparing our results with those reported by Štajner and Mitkov (2011), we will also be able to examine whether the religious text had followed the same trends of diachronic stylistic changes as the broader text category they belong to (Prose).

3. Methodology

As some of the corpora were not publicly available in their tagged versions, we decided to use the raw text version of all corpora and parse it with the state-of-the-art Connexor's Machinese syntax parser[5], following the methodology for feature extraction proposed by Štajner and Mitkov (2011). We agree that this approach allows us to have a fairer comparison of the results among different corpora and to achieve a more consistent, highly accurate sentence splitting, tokenisation, lemmatisation and part-of-speech tagging. As the details of tokenisation and lemmatisation process of this parser (Connexor's Machinese syntax parser) were already discussed in detail by Štajner and Mitkov (2011), the focus in this study will be on the POS tagging.

3.1. Part-of-speech tagging

Connexor's Machinese Syntax parser reported the POS accuracy of 99.3% on Standard Written English (benchmark from the Maastricht Treaty) and there was no ambiguity Connexor (2006). For each known word the parser assigns one of the 16 possible morphological (POS) tags: N (noun), ABBR (abbreviation), A (adjective), NUM (number), PRON (pronoun), DET (determiner), ADV (adverb), ING (present participle), EN (past participle), V (verb), INTERJ (interjection), CC (coordinative conjunction), CS (subordinate conjunction), PREP (preposition), NEG-PART (negation particle *not*), INFMARK (infinitive marker *to*).

Here it is important to note that Connexor's Machinese parser differentiate between present and past participle (ING and EN), and verbs (V). This should be taken into account later in Section 4, where diachronic changes of the POS frequencies are presented and discussed. It is also important to emphasise that in the newer version of the parser, the EN and ING forms, which can represent either present and past participle or corresponding nouns and adjectives, are assigned a POS tag (EN, ING, N or A) according to their usage in that particular case. For example, in the sentence:

> *"Some of the problems were reviewed yesterday at a meeting in Paris..."* (LOB:A02),

the word *meeting* was assigned the N tag, while in the sentence:

> *"... Mr. Pearson excels in meeting people informally... "* (LOB:A03),

the same word *meeting* was assigned the ING tag. Similarly, the word *selected* was assigned the A tag in the sentence:

> *"The editors ask some 20 to 30 working scientists to report on the progress made in selected and limited fields... "* (LOB:C14),

[4] http://www.ranks.nl/resources/stopwords.html

[5] www.connexor.eu

while in the other sentence:

> "... *Miss Anna Kerima was selected as best actress...* " (LOB:C02),

the same word *selected* was assigned the EN tag.

3.2. Feature Extraction

All features were separately calculated for each text in order to enable the use of statistical tests of differences in means (Section 3.3) The first four features (ASL, ARI, LD, LR) were computed using the formulas given in Section 1.2. The second set of features (POS frequencies) were calculated separately for each POS tag and for each text, as the total number of that specific POS tag divided by the total number of tokens in the given text (eq.5).

$$< POS >= \frac{total\_number\_of\_ < POS >}{total\_number\_of\_tokens} \tag{5}$$

Stop words were calculated in a similar way. For each stop word and each text, the corresponding feature was calculated as the total number of repetitions of that specific stop word divided by the total number of tokens for the given text (eq.6).

$$< STOP >= \frac{total\_number\_of\_ < STOP >}{total\_number\_of\_tokens} \tag{6}$$

3.3. Experimental Settings

We conducted two sets experiments:

- Diachronic changes of style in British English (1961–1991)
- Diachronic changes of style in American English (1961–1992)

For each experiment we calculated the statistical significance of the mean differences between the two corresponding groups of texts for each of the features.

Statistical significance tests are divided into two main groups: parametric (which assume that the samples are normally distributed) and non-parametric (which does not make any assumptions about the sample distribution). In the cases where both samples follow the normal distribution, it is recommended to use parametric tests as they have greater power than the non-parametric ones. Therefore, we first applied the Shapiro-Wilk's W test (Garson, 2012a) offered by SPSS EXAMINE module in order to examine in which cases/genres the features were normally distributed. This test is a standard test for normality, recommended for small samples. It shows the correlation between the given data and their expected normal distribution scores. If the result of the W test is 1, it means that the distribution of the data is perfectly normal. Significantly lower values of W (≤ 0.05) indicate that the assumption of normality is not met.

Following the discussion in (Garson, 2012b), in both experiments we used the following strategy: if the two data sets we wanted to compare were both normally distributed we used the t-test for the comparison of their means; if at least one of the two data sets was not normally distributed, we used the Kolmogorov-Smirnov Z test for calculating the statistical significance of the differences between their means. Both tests were used in their two independent sample versions and the reported significance was the two-tailed significance. After applying the statistical tests, we only focused on the features which demonstrated statistically significant change (at a 0.05 level of significance) in the observed period. We presented and discussed those changes in Sections 4.1 and 4.2.

After that, we applied several machine learning classification algorithms in Weka[6](Hall et al., 2009; Ian H. Witten, 2005) in order to see which group of the features would be the most relevant for diachronic classification of religious texts. In order to do so, we first applied two well-known classification algorithms: Support Vector Machines (Platt, 1998; Keerthi et al., 2001) and Naive Bayes (John and Langley, 1995) to classify the texts according to the year of publication (1961 or 1991/2), using all features which reported a statistically significant change in that period. The SVM (SMO in Weka) classifier was used with two different settings. The first version used previously normalised features and the second – previously standardised features. Furthermore, we tried the same classification using all possible combinations of the three sets of features: only the first set (1), only the second set (2), only the third set (3), the first and second set together (1+2), the second and third set (2+3), the first and third set (1+3). Then we compared these classification performances with the ones obtained by using all three sets of features together (1+2+3) in order to examine which features are the most relevant for the diachronic classification of this text genre. The results of these experiments are presented and discussed in Section 4.3.

4. Results and Discussion

The results of the investigation of diachronic stylistic changes in religious texts are given separately for British and American English in the following two subsections and compared in the second subsection. The results of the machine learning classification algorithms for both language varieties and the discussion about the most relevant feature set for diachronic classification are given in the third subsection.

4.1. British English

The results of diachronic changes in religious texts written in British English are given in Table 3. The table contains information only about the features which reported a statistically significant change (sign. ≤ 0.05). The columns '1961' and '1991' contain the arithmetic means of the corresponding feature in 1961 and 1991, respectively. The column 'Sign.' contains the p-value of the applied t-test or, alternatively, the p-value of Kolmogorov-Smirnov Z test (denoted with an '*') for the cases in which the feature was not normally distributed in at least one of the two years (according to the results of the previously applied Shapiro-Wilk's W test as discussed in Section 3.3). Column 'Change' contains the relative change calculated as a percentage of the

[6] http://www.cs.waikato.ac.nz/ml/weka/

feature's starting value in 1961. The sign '+' stands for an increase and the sign '−' for a decrease over the observed period. In case the frequency of the feature was '0' in 1961 and different than '0' in 1991, this column contains value 'NA' (e.g. feature 'whom' in Table 3).

Feature	1961	Sign.	Change	1991
ARI	10.332	0.010	+37.72%	14.229
LD	0.304	0.027	+8.47%	0.329
LR	0.268	0.030	+9.40%	0.293
V	14.087	0.020	−14.32%	12.070
PREP	12.277	0.016	+11.69%	13.712
A	6.843	0.009	+24.37%	8.511
ING	0.957	0.042	+31.38%	1.257
an	0.237	0.008	+52.27%	0.361
as	0.554	0.034	+32.68%	0.736
before	0.090	0.002*	−79.98%	0.018
between	0.062	0.046*	+112.07%	0.132
in	1.781	0.046*	+30.57%	2.325
no	0.230	0.027	−46.37%	0.123
these	0.191	0.017*	−45.70%	0.104
under	0.059	0.046*	−77.80%	0.013
whom	0.000	0.006*	NA	0.044
why	0.054	0.046*	−71.25%	0.015

Table 3: British English (1961–1991)

The increase of ARI (Table 3) indicates that religious texts in British English were more complex (in terms of the sentence and word length) and more difficult to understand in 1991 than in 1961. While in 1961, these texts required an US grade level 10 on average for their comprehension, in 1991, they required an US grade level 14. Also, the increase of LD and LR in this period (Table 3) indicates the usage of much wider and more diverse vocabulary in these texts in 1991 than in 1961.

The results also demonstrated changes in the frequency of certain word types during the observed period. Verbs (excluding the past and present participle forms) were used less in 1991 than in 1961, while the prepositions, adjectives and present participles were more frequent in religious texts written in 1991 than in those written in 1961 (Table 3).

The frequency of certain stop words in religious texts had also significantly changed over the observed period (1961–1991). The most striking is the change in the use of the word 'between' which was used more than twice as much in texts written in 1991 than in those written in 1961. Whether this was the consequence of an increased use of some specific expressions and phrases containing this word, remains to be further investigated. Also, it is interesting to note that the word 'whom' was used not even once in the texts from 1961 while it was considerably often used in the texts from 1991.

4.2. American English

The results of the investigation of diachronic stylistic changes in religious texts written in American English are given in Table 4, using the same notation as in the case of British English.

The first striking difference between diachronic changes reported in British English (Table 3) and those reported in American English (Table 4) is that in American English none of the four features of the first set (ASL, ARI, LD, LR) demonstrated a statistically significant change in the observed period (1961–1992), while in British English three of those four features did. Actually, the only feature which reported a significant change in both language varieties during the period 1961–1991/2 is the frequency of adjectives (A), which had increased over the observed period in both cases. This might be interpreted as a possible example of Americanisation – "the influence of north American habits of expression and behaviour on the UK (and other nations)" (Leech, 2004), in the genre of religious texts.

Feature	1961	Sign.	Change	1991
N	27.096	0.009	+10.72%	30.002
PRON	7.763	0.046*	−30.04%	5.431
A	8.092	0.040	+19.63%	9.680
ADV	6.050	0.020	−14.58%	5.168
all	0.298	0.017*	−36.99%	0.188
have	0.500	0.010	−36.52%	0.317
him	0.240	0.017*	−86.25%	0.033
it	0.871	0.015	−32.79%	0.586
not	0.661	0.046*	−13.37%	0.572
there	0.152	0.017*	−42.42%	0.088
until	0.052	0.017*	−60.14%	0.021
what	0.237	0.046*	−9.26%	0.215
which	0.054	0.046*	−48.63%	0.028

Table 4: American English (1961–1992)

On the basis of the results presented in Table 4, we observed several interesting phenomena of diachronic changes in American English. The results reported a significant increase of noun and adjective frequency, and a significant decrease of pronoun frequency. These findings are not surprising, given that adjectives usually play the function of noun modifiers (Biber et al., 1999) and therefore, an increase of noun frequency is expected to be followed by an increase of adjective frequency. Also, as pronouns and full noun phrases usually compete for the same syntactic positions of subject, object and prepositional complement (Hudson, 1994; Biber et al., 1999), a noun increase and pronoun decrease are not unexpected to be reported together (Mair et al., 2003).

4.3. Feature Analysis

As it was discussed previously in Section 3.3, we used machine learning classification algorithms in order to examine which set of features would be the most important for diachronic classification of religious texts in 20th century. We tried to classify the texts according to the year of publication (1961 and 1991 in the case of British English, and 1961 and 1992 in the case of American English), using all possible combinations of these three sets of features: (1), (2), (3), (1)+(2), (1)+(3), (2)+(3), and compare them with the classification performances when all three sets of features are used (1)+(2)+(3). The main idea is that if a set of features is particularly important for the classification, the performance of the classification algorithms should significantly drop when this set of features is excluded. All experiments were conducted using the 5-fold cross-validation with 10 repetitions in Weka Experimenter.

The results of these experiments for British English are presented in Table 5. Column 'Set' denotes the sets of features used in the corresponding experiment. Columns 'SMO(n)', 'SMO(s)' and 'NB' stand for the used classification algorithms – Support Vector Machines (normalised), Support Vector Machines (standardised) and Naive Bayes, respectively. The results (classification performances) of each experiment were compared with the experiment in which all features were used (row 'all' in Table 5), using the two-tailed paired t-test at a 0.05 level of significance provided by Weka Experimenter. Cases in which the differences between classifier performance of the particular experiment was significantly lower than in the experiment where all features were used are denoted by an '*'. There were no cases in which a classifier's accuracy in any experiment outperformed the accuracy of the same classifier in the experiment with all features.

Set	SMO(n)	SMO(s)	NB
all	90.19	92.52	85.48
(1)	68.81*	64.05*	67.86*
(2)	68.19*	70.14*	70.33*
(3)	87.24	92.67	88.81
(2)+(3)	91.14	93.33	86.48
(1)+(3)	87.38	89.52	88.43
(1)+(2)	74.48*	66.71*	70.33*

Table 5: Diachronic classification in British English

From the results presented in Table 5 we can conclude that in British English, the changes in the frequencies of the stop words (third set of features) were the most important for this classification task. All experiments which did not use the third set of features (rows '(1)', '(2)' and '(1)+(2)' in Table 5), reported a significantly lower performances of all three classification algorithms.

In the case of American English, we needed to compare only the results of the experiment in which all features were used (row 'all' in Table 6) with those which used only the second (POS frequencies) or the third (stop-words) set of features, as none of the features from the first set (ASL, ARI, LD and LR) had demonstrated a significant change in the observed period 1961–1992 (Table 4, Section 4.2). The results of these experiments are presented in Table 6.

Set	SMO(n)	SMO(s)	NB
all	73.67	77.86	72.90
(2)	70.57	67.67	71.10
(3)	74.67	76.24	78.33

Table 6: Diachronic classification in American English

In diachronic classification of religious texts in American English, no significant difference was reported in the performance of the classification algorithms between the experiment in which both sets of features (POS frequencies and stop-words) were used and those experiments in which only one set of features (either POS frequencies or stop-words) was used. Therefore, based on the results of these experiments (Table 6) we were not able to give a priority to any of these two sets of features in the diachronic classification task.

The comparison of the results of diachronic classification between British and American English (Table 5 and Table 6) lead to the conclusion that the stylistic changes in religious texts were more prominent in British than in American English, as all three classification algorithms in British English (row 'all' in Table 5) outperformed those in American English (row 'all' in Table 6). This conclusion is also in concordance with the comparison between British and American diachronic changes based on the relative changes of investigated features reported in Tables 3 and 4 (Sections 4.1 and 4.2).

5. Conclusions

The presented study offered a systematic and NLP oriented approach to the investigation of style in 20th century religious texts. Stylistic features were divided into three main groups: (ASL, ARI, LD, LR), POS frequencies and stop-words frequencies.

The analysis of diachronic changes in British English in the period 1961–1991 demonstrated significant changes of many of these features over the observed period. The reported increase of ARI indicated that religious texts became more complex (in terms of average sentence and word length) and more difficult to read, requiring a higher level of literacy and education. At the same time, the increase of LD and LR indicated that the vocabulary of these texts became wider and richer over the observed period (1961–1991). The investigation of the POS frequencies also demonstrated significant changes, thus indicating that 30 years time gap is wide enough for some of these changes to be noticed. The results reported a decrease in verb frequencies (excluding the present and past participles) and an increase in the use of present participles, adjectives and prepositions. The analysis of the stop-words frequencies indicated a significant changes in the frequency of ten stop-words (an, as, before, between, in, no, these, under, whom, why) with the most prominent change in the case of the word 'between'.

The results of the machine learning experiments pointed out the third set of features (stop-words frequency) as the most dominant/relevant set of features in the diachronic classification of religious texts in British English. These results were in concordance with the results of the statistical tests of mean differences which reported the highest relative changes exactly in this set of features.

The investigation of stylistic features in 20th century religious texts in American English reported no significant changes in any of the four features of the first set (ASL, ARI, LD, LR) in the observed period 1961–1992. The analysis of the second set of features (POS frequencies) indicated an increase in the use of nouns and adjectives, and a decrease of pronoun and adverb frequencies. In the third set of features (stop-words frequencies), nine words reported a significant decrease in their frequencies (all, have, him, it, not, there, until, what, which). The machine learning experiments, which had the aim of pointing out the most relevant set of stylistic features for diachronic classification of religious texts, did not give the preference to any of the two sets of features (POS and stop-words frequencies) in this task.

The comparison of diachronic changes in the period 1961–1991/2 between British and American English indicated very different trends of stylistic changes in these two language varieties. From all 146 investigated features, only one feature (adjective frequency) reported a significant change (increase) in both – British and American English, during the observed period (1961–1991/2). The overall relative change was much higher in British than in American English, which was additionally confirmed by the results of the machine learning classification algorithms.

6. References

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Connexor, 2006. *Machinese Language Analysers. Connexor Manual.*

Gloria Corpas Pastor, Ruslan Mitkov, Afzal Naveed, and Pekar Victor. 2008. Translation universals: Do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the AMTA*.

Nelson W. Francis. 1965. A standard corpus of edited present-day American english. *College English*, 26(4):267–273.

Michael Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

David G. Garson. 2012a. Testing of assumptions: Normality. Statnotes: Topics in Multivariate Analysis.

David G. Garson. 2012b. Tests for two independent samples: Mann-Whitney U, Wald-Wolfowitz runs, Kolmogorov-Smirnov Z, & Moses extreme reactions tests. Statnotes: Topics in Multivariate Analysis.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November.

David Holmes. 1994. Authorship attribution. *Computers and the Humanities*, 28:87–106. 10.1007/BF01830689.

David I. Holmes. 1998. The evolution of stylometry in humanities scholarship.

Richard Hudson. 1994. About 37% of word-tokens are nouns. *Language*, 70(2):331–339.

Eibe Frank Ian H. Witten. 2005. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

George H. John and Pat Langley. 1995. Estimating continuous distributions in Bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13(3):637–649.

Peter J. Kincaid and Leroy J. Delionbach. 1973. Validation of the Automated Readability Index: A follow-up. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 15(1):17–20.

Geoffrey Leech and Nicholas Smith. 2005. Extending the possibilities of corpus-based research on English in the twentieth century: A prequel to LOB and FLOB. *ICAME Journal*, 29:83–98.

Geoffrey Leech and Nicholas Smith. 2009. Change and constancy in linguistic change: How grammatical usage in written English evolved in the period 1931-1991. *Language and Computers*, 69(1):173–200.

Geoffrey Leech. 2004. Recent grammatical change in English: data, description, theory. *Language and Computers*, 49(1):61–81.

Christian Mair, Marianne Hundt, Geoffrey Leech, and Nicholas Smith. 2003. Short term diachronic shifts in part of speech frequencies: A comparison of the tagged LOB and FLOB corpora. *International Journal of Corpus Linguistics*, 7(2):245–264.

Douglas R. McCallum and James L. Peterson. 1982. Computer-based readability indexes. In *Proceedings of the ACM '82 conference*, ACM '82, pages 44–48, New York, NY, USA. ACM.

John C. Platt. 1998. Fast training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning*, 208:1–21.

R. J. Senter and E. A. Smith. 1967. Automated Readability Index. Technical report, Defense Technical Information Center. United States.

J. Smith and C. Kelly. 2002. Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities*, 36:411–430. 10.1023/A:1020201615753.

Louis Ule. 1982. Recent progress in computer methods of authorship determination. *Association of Literary and Linguistic Computing Bulletin*, 23(2):73–89.

Sanja Štajner and Ruslan Mitkov. 2011. Diachronic stylistic changes in British and American varieties of 20th century written English language. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage at RANLP 2011*, pages 78–85.

Multi-Word Expressions in the Spanish Bhagavad Gita, Extracted with Local Grammars Based on Semantic Classes

Daniel Stein
Hamburg Centre for Language Corpora, University of Hamburg
Max-Brauer-Allee 20, 22765 Hamburg, Germany
ds@daniel-stein.com

Abstract

As religious texts are often composed in metaphorical and lyrical language, the role of multi-word expressions (MWE) can be considered even more important than usually for automatic processing. Therefore a method of extracting MWE is needed, that is capable of dealing with this complexity. Because of its ability to model various linguistic phenomena with the help of syntactical and lexical context, the approach of Local Grammars by Maurice Gross seems promising in this regard. For the study described in this paper I evaluate the use of this method on the basis of a Spanish version of the Hindu poem Bhagavad Gita. The search will be refined on nominal MWE, i.e. nominal compounds and frozen expressions with two or three main elements. Furthermore, the analysis is based on a set of semantic classes for abstract nouns, especially on the semantical class "phenomenon". In this article, the theory and application of Local Grammars is described, and the very first results are discussed in detail.

Keywords: Religious Texts, Multiword-Expression, Local Grammars

1. Introduction

The *Bhagavad Gita*[1] (short: Gita) is one of the central spiritual texts of Hinduism. It is considered to be a synthesis of several religious and philosophical schools of ancient India, including the *Upanishads*, the *Vedas*, and *Yoga*. In Hinduism, a distinction between revealed (*Śruti*) and remembered (*Smriti*) sacred texts is common. But the classification of the Gita depends on the relative branch of Hinduism, as it is a part of a collection of remembered texts as well as a revelation of the god *Krishna*.

In a greater context, the Gita represents the chapters 25-42 of the epic poem Mahabharata ("Great India"). Vyasa, the compiler of the poem comes also into question as the author. The content deals mainly with the conversation between the embodied god Krishna and the prince *Arjuna* on a battlefield inside a chariot. Arunja becomes doubtful when he gets aware that members of his family are fighting for the opposite side. Lord Krishna uses philosophical and spiritual arguments to induce Arjuna to face this war nevertheless. For many commentators, this war stands as an allegory for life itself or the human weaknesses one has to fight, but literal interpretations are also present in the commentaries.

2. Linguistic Analysis

From a linguistic point of view, the Bhagavad Gita is a poem composed of 700 verses in 18 chapters. As this work is a first case study for my dissertation project (that deals with Spanish MWE in various domain texts), I use a Spanish version of the Gita for this analysis. It is published online by the *International Society of Krishna Consciousness* (ISKCON) [2] under the name "El Bhagavad-gita Tal Como Es".

As supposed to be common religious texts or poems, also the Gita is a tight and profoundly complex work with very metaphorical language. Due to the fact that it is a junction of different religious systems, the content of the Gita sometimes seems to be contradictory, e.g. in relation to the question of duality or unitary of being.

Another problem (not only for the analysis) arising for religious texts in general is the questionable possibility of using translations as source: As a matter of fact, a translation always contains an interpretation and (more or less slight) loss of information. Furthermore, a given language's metaphorical character is based on the lexical inventory of this language. So a translation can make it impossible to use the same metaphorical images. This is especially the case for translations of lyrical texts in which not only semantic information needs to be translated, but also measure and rhyme may be essential parts of understanding. In religious texts this may result in serious difficulties. Nevertheless many religious communities accept translated versions of their sacred texts (at least to offer them in a first place to other cultures) and so does Hinduism, at least those branches that offer translations to other languages.[3]

For a non-Hindu reader a further problem is the large number of proper names and even whole semantic concepts with ancient Indian background that are not necessarily translatable to languages with another cultural background (e.g. "*Dharma*" which can roughly be translated as "*that which upholds, supports or maintains the regulatory order of the universe*" [4] and it becomes

[1] In Sanskrit भगवद्गीता "Song of God"

[2] The Spanish version is based on the English translation "The Bhagavad - Gītā As It Is" by Abhay Charan Bhaktivedanta Swami Prabhupada, the founder of the ISKCON. It was first published in 1968 and was the foundation for translations into about 60 languages. Next to the translated text it contains the Sanskrit original, a transliteration and extensive commentary. http://www.spiritual-revolutionary.com/Espanol/BG_espanol/BG_Indice.htm (Last visit: 01/26/2012)

[3] Although this very translation by the ISKCON is one of the widest spread of the Gita, it is still not uncriticized, cf. http://www.dvaita.org/shaastra/gita/prabhupada_review.shtml (Last visit: 26/01/2012).

[4] According to http://en.wikipedia.org/wiki/Dharma (Last visit: 01/24/12).

even more complex if one considers the different meaning in the context of Buddhism).

3. MWE in Religious Texts

In Natural Language Processing, there are several methods and tools that may be useful for the study of religious texts like the Bhagavad Gita and their translations. This includes methods for an automatic alignment of the original source with translated versions, or semantic search engines (e.g. Shazadi, 2011).

But in order to get these methods to work, a text analysis that copes with the complexity of these metaphorical rich and semantically complex texts is necessary. In this specific context, multi-word expressions presumably may play an (even more?) vital role as they already do in nearly every part of NLP. The idiomatic structure and the high level of information density in this text form is presumably very rich in MWE. This information needs to be known for satisfying automatic analyses.

The term MWE regularly subsumes a broad scope of linguistic phenomena [5], so it is useful to choose a well-defined part for the first research.

A categorization of MW can be based on one or more different features, e.g. the word forms, number or syntactical structure of the main elements, the distribution (collocations) etc.

For this paper I'm going to follow the categorization by Guenthner and Blanco Escoda (2004). I'm focussing on nominal MWE which includes the types Compound Nouns (e.g. *agujero negro* = black hole) and Frozen Modifiers (e.g. *miedo cerval* = terrible fear).

A further useful semantical distinction can be made between nouns that describe facts (e.g. *pecado mortal* = deadly sin), and nouns that describe entities (e.g. *obispo auxiliar* = auxiliary bishop).

These categories have been classified in detail for Spanish in the context of the *Nuevo diccionario histórico de la lengua española* (New Historical Dictionary of the Spanish Language) by (Blanco Escoda, 2010). Research has proven semantic categories may vary between languages, so it is recommendable to use a classification that was created for Spanish. I use the semantical class of facts (Spanish: *hechos*). This classification is named *Spanish Hierarchy of Semantic Labels of Facts* (SHSL_F). The MWE that belong into this category are supposed to be easier to identify via context (regarding e.g. the use of auxiliary verbs) than those of the semantical class of entities. The SHSL_F is divided into the following 17 categories:

- Acción (Action)
- Acontecimiento (Event)
- Actitud (Attitude)
- Actividad (Activity)
- Cantidad (Quantity)
- Característica (Characteristics)
- Comportamiento (Behavior)
- Conjunto de Hechos (Set of facts)
- Costumbre (Habit)
- Estado (State)
- Fenómeno (Phenomenon)
- Grado (Degree)
- Parámetro (Parameter)
- Período (Period)
- Proceso (Process)
- Relación Factual (Factual Relationship)
- Situación (Situation)

For the study in this paper I focus on phenomena that may be perceived via the sensual organs or that can be understood. A complete graph of the semantic class of phenomena would include also other aspects e.g. physiological phenomena like a pulse.

4. Tools

4.1 Unitex

The software that is used for this analysis is Unitex 2.1[6] by Sébastien Paumier. Unitex is an open source corpus processing tool that allows the construction of so called Local Grammars (Gross, 1997). This formalism is used to model complex linguistic phenomena using syntactical and lexical context and is often visualized as directional graphs, although technically it is a recursive transition network (cf. figure 1). Local Grammars are useful for many issues in NLP, such as lexical disambiguation, representation of lexical variations in dictionaries or information extraction. They also are very useful for the identification of MWE.

4.2 DELA

The use of Local Grammars relies heavily on a special kind of dictionary, called DELA (Dictionnaires Electroniques du LADL[7]). In DELA, MWE are treated exactly the same way as simple lexical units. The structure of a DELA lemma is as follows:

inflected form , canonical form . syntactical code + semantical code : inflectional code / comment

apples,apple.N+conc:p/this is a comment

With the original installation of Unitex, a basic dictionary of Spanish is included that was designed by the fLexSem group from the Universitat Autónoma de Barcelona (Blanco Escoda, 2001). This basic dictionary contains 638,000 simple words. Considering the inflectional character of Spanish, this is a moderate size which is reflected in the lack of even some common words (see below). Nevertheless the basic dictionary is a good starting point for research and is used for this analysis.

[5] This reflects in the high number of denominations used to describe them, e.g. Collocations, Frozen Expressions, Terms or Compounds, to mention just a few.

[6] http://igm.univ-mlv.fr/~unitex/ (Last visit: 01/29/2012), also for an overview of studies based on Unitex.

[7] Laboratoire d'Automatique Documentaire et Linguistique, the institution in Paris where Maurice Gross developed Local Grammars.

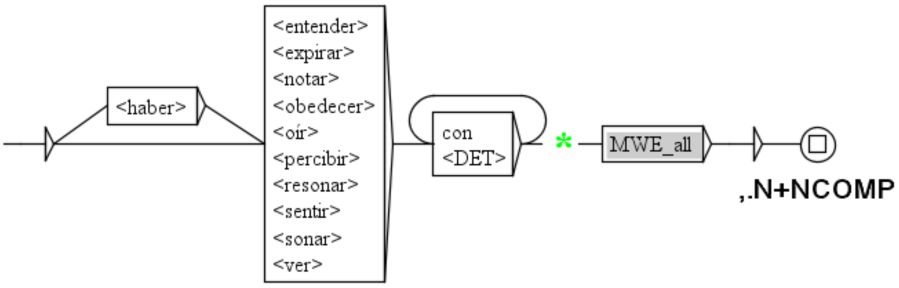


Figure 1: Early version of the Local Grammar for the semantic class "Phenomenon".

Additionally it will be supplemented by some self-constructed dictionaries based on the special features of the Gita.

To make the Gita a useable corpus and to create the mentioned dictionaries, the text needs to be stripped of comments and titles. Also the letter š (which is relevant for the Sanskrit transliterations) has to be added to the Spanish alphabet file in the according Unitex directory. After this step the text has to be loaded as a corpus into Unitex in order to get an overview of its lexical inventory. This can be divided into three lists: The simple and the complex units (that are found in the dictionary as well as in the corpus) and a list of units that just appear in the corpus (and are considered as unknown).

In the case of the Gita, there are 4437 simple-word lexical entries, 0 compounds and 392 unknown forms. 222 items of the unknown forms are regular Spanish words that are not contained in the basic dictionary (e.g. *adelantada* (advanced), *confundida* (confused) but also *yoga*[8]), the other 170 items consist of a special vocabulary of transliterated Sanskrit words which are mostly names (as *Arunja*) with a few exceptions (as the "sacred syllable" *óm*). It is reasonable to create two dictionaries out of this list, one for the Spanish words that are unknown to the basic dictionary to make them useable for further investigations and one just for the Sanskrit words limited for the use with the Gita or similar texts as a corpus.

Corpus processing with Unitex works best with no unknown words so this is an essential issue and easy to accomplish for a short text like the Gita. After applying the newly generated dictionaries on the Corpus in Unitex, it possesses the features presented in table 1.

	Bhagavad Gita
Tokens	53341
Simple word lexical entries	4829
Compound lexical entries	0
Unknown simple words	0

Table 1: Lexical Units in "El Bhagavad-gita Tal Como Es"

4.3 Local Grammars

The functionality of Local Grammars as a way to extract MWE will be explained using the example graph in figure 1 and the corresponding outputs in figure 2. In simple terms, the Local Grammar in figure 1 matches a phrase, that may begin with an inflected form of the word *haber* (which is part of a perfect tense construction) or with an inflected form of one of the verbs in the list, followed by the word *con* or a determiner or both and at least it needs to find its way through a set of graphs called MWE_all.

Example phrase and recognized path[9]:

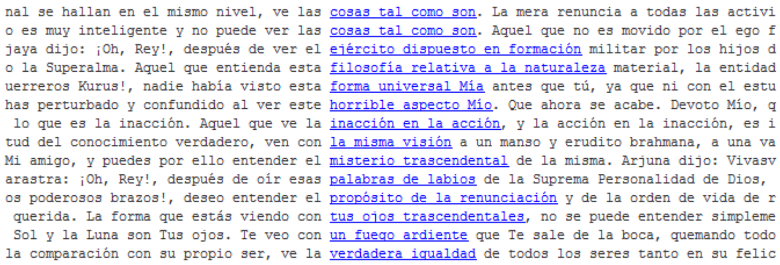
He	*entendio*	*el*	*agujero*	*negro*
<haber>	**<entender>**	**<DET>**	**N**	**A**

The graph shown in figure 1 is relatively simple but sufficient to explain the basic principles of Local Grammars and will now be analysed in detail:

- The arrow on the left side represents the starting point; the square within a circle on the right is the end point.
- A text analysis results in a concordance of a number of passages of text that matches to the Local Grammar. A match is achieved when a text passage is found, that is able to be reconstructed by a valid connection from the start point to the end point of the Local Grammar.
- Words without brackets match in cases where exactly the same form is found in the corpus, e.g. *con* (with).
- Words in angle brackets match all inflected forms of this word. E.g. the expression <*entender*> (to understand) matches *entender*, *entiendo*, *entendemos* etc.
- Semantic or syntactic codes in brackets as <DET> represent all words with the same code in the lexicon (here: DET = determiner, e.g. *el*, *la*, *esto*… (he, she, this…).
- Loops are also possible to deal with recursion or some kinds of variation (e.g. this loop matches for *con*, *el*, *con el*, *el con la;* etc:).

[8] Additionally I corrected the misspelled words I found in the Gita in order to improve the analysis, e.g. ahbía -> *había*.

[9] N = Noun, A = Adjective



```
nal se hallan en el mismo nivel, ve las cosas tal como son. La mera renuncia a todas las activi
o es muy inteligente y no puede ver las cosas tal como son. Aquel que no es movido por el ego f
jaya dijo: ¡Oh, Rey!, después de ver el ejército dispuesto en formación militar por los hijos d
o la Superalma. Aquel que entienda esta filosofía relativa a la naturaleza material, la entidad
uerreros Kurus!, nadie había visto esta forma universal Mía antes que tú, ya que ni con el estu
has perturbado y confundido al ver este horrible aspecto Mío. Que ahora se acabe. Devoto Mío, q
 lo que es la inacción. Aquel que ve la inacción en la acción, y la acción en la inacción, es i
tud del conocimiento verdadero, ven con la misma visión a un manso y erudito brahmana, a una va
Mi amigo, y puedes por ello entender el misterio trascendental de la misma. Arjuna dijo: Vivasv
arastra: ¡Oh, Rey!, después de oír esas palabras de labios de la Suprema Personalidad de Dios,
os poderosos brazos!, deseo entender el propósito de la renunciación y de la orden de vida de r
 querida. La forma que estás viendo con tus ojos trascendentales, no se puede entender simpleme
 Sol y la Luna son Tus ojos. Te veo con un fuego ardiente que Te sale de la boca, quemando todo
la comparación con su propio ser, ve la verdadera igualdad de todos los seres tanto en su felic
```

Figure 2: Concordance for the graph in figure 1.

- The grey shaded box is a link to another graph, in this case to a set of sub graphs that need to be matched, too, in order to achieve a match in the whole graph. The linked grammars are a set that represents all potential possible syntactical variations [10] of MWE in Spanish. The combination of contextual and syntactical Local Grammars is essential for this approach. One can say that the graph in figure 1 is the context that allows the MWE_all main graph to match with higher precision. To get an overview of all the graphs involved in a complete analysis see figure 3.
- It is possible to define left and/or right context that will not be part of the output. The asterisk defines the end of the left context, so everything to the right side of the asterisk is considered as output, i.e. the MWE itself.
- The graph also can be used as a transducer that inserts text into the output. The bold black text beneath the arrow on the right is DELA encoding and can be attached automatically to the output e.g. to create a DELA dictionary of the found compound forms.

As the graph is still in development, this version is considered as a rough sketch and far from the complexity a Local Grammar can reach. After the application of the graph to the "Bhagavad Gita – Tal Como Es" corpus, the 14 passages listed in the concordance in figure 2 are the result. Via application of all the graphs that are not yet involved in the process (cf. figure 3), a significantly larger number of results can be expected.

5. MWE from Bhagavad Gita

The results displayed in figure 2 are of a surprisingly high quality, at least considering the simplicity of the underlying Local Grammar and the complexity of the corpus. When examined in detail, there are still many errors in the expressions found. This implies that the approach of using auxiliary verbs for a Local Grammar to extract MWE of facts is useful. Although a semantic analysis of the phrases obtained would be of interest e.g. in regards to automatic tagging, it is not in the scope of this paper. So I will only analyse whether the expressions in the concordance are MWE and of what syntactical structure they are. I will use a broad definition of MWE including those that are idiomatic as well as those that are common but not obligatory (or fixed) combinations. A helpful rule of thumb is the question if a group of words could possibly be expressed using a single word in another language. An analysis of the phrases reveals the following:

1. ~~*cosas tal como son*~~ (the things as they are): No nominal MWE. The verb *son* is recognized as a

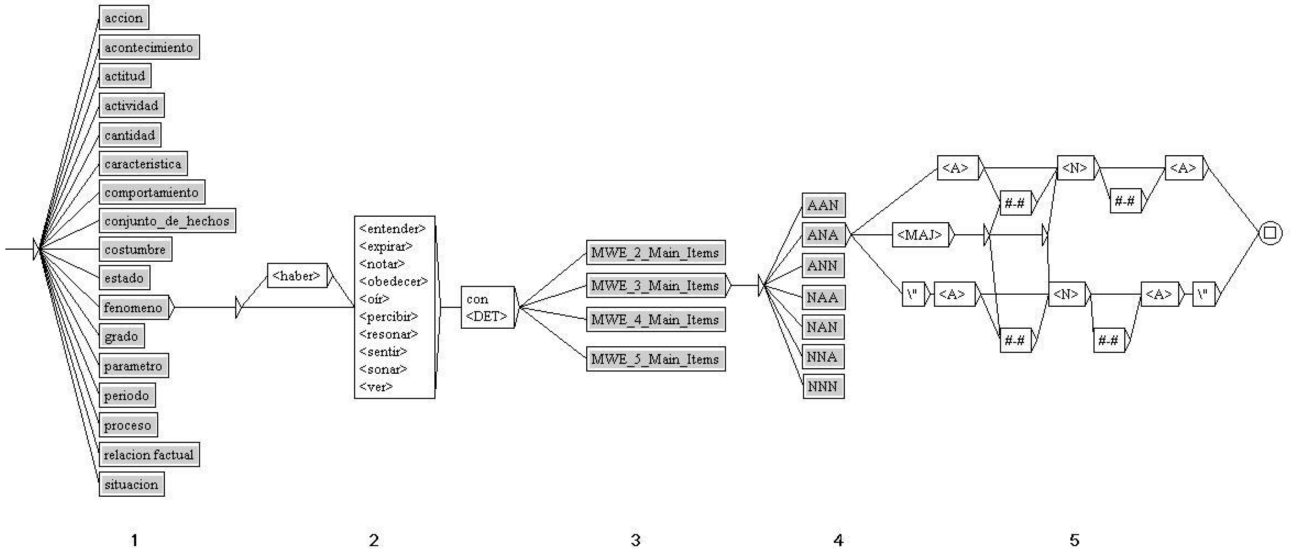


Figure 3: Exemplified combination of the linked Local Grammars, (1) the semantic classes, (2) the class "fenomeno", (3) the MWE overview, (4) MWE with three main items and (5) the MWE with ANA structure.

[10] In fact, for this study I limit the possible MWE for such with two or three main elements.

noun. Both interpretations are present in the dictionary because both may be correct in the right context. But in this case it is unwanted and can be corrected by changing the lexicon entry. As it is impossible to change the standard dictionaries of Unitex, it is necessary to create a lexicon which overrides the standard one. This can be achieved by adding a minus at the end of the dictionary's name (e.g. wronggita-.dic) in order to assign a higher priority to it. For details see the Unitex manual (Paumier, 2010). For this study such a dictionary was prepared containing the four most typical problems: the words *a*, *de*, *la* and *y* encoded as nouns.

2. *~~cosas tal como son~~*: cf. 1
3. *ejercito dispuesto en formación* [*militar*] (army positioned in military formation): NANA – but because of the exclusion of MWE with four basic elements it is not recognized.
4. *filosofía relativa a la naturaleza* [*material*] (philosophy concerning the material nature): NANN, cf. 3.
5. *Forma universal ~~Mía~~* (universal form of myself): two basic elements, NA – *Mía* is not a noun and needs to be changed in the dictionary, too. However, in this special case Mía is capitalized as it is used by Krishna to refer to himself. The capitalization of pronouns may be a special characteristic of religious texts which has to be taken into account in automatic processing.
6. *horible aspecto ~~Mío~~* (horrible aspect of myself): NA, *Mío* also isn't a noun but capitalized for the self-reference by Krishna, too, cf. 5.
7. *~~un~~ fuego ardiente* (a burning fire): NA, but the indefinite article un needs to be corrected in the Local Grammar.
8. *inacción en la acción* (inaction in action): NN.
9. *~~la~~ misma vision* (same vision): AN, but the defined article la needs to be corrected in the Local Grammar.
10. *misterio transcendental* (transcendental mystery): NA.
11. *palabras de labios* (vain words): NN.
12. *propósito de la renunciación* (intention to renunciation): NN.
13. *~~tus~~ ojos transcendentales* (your transcendental eyes), NA, but the possessive pronoun *tus* needs to be corrected in the Local Grammar.
14. *verdadera igualdad* (true equality): AN.

12 out of the 14 expressions actually contain nominal MWEs. The majority of problems can be corrected by adjusting the dictionary in order to prevent unwanted interpretations of ambiguous words to be applied. Two errors are due to the Local Grammars and can be corrected there. From this point of development work can still be done to improve the recognition of the MWE and to refine the Local Grammars for the context of the Gita. Using the same graphs for other texts should also result in useful data, but dictionaries as well as Local Grammars may need adaption. A more comprehensive evaluation of this approach is beyond the scope of this paper. It would require more advanced graphs for more semantical classes in order to be comparable to other MWE identification approaches as e.g. statistical based ones. Due to the early stage of my study, I was not yet able to realize this.

6. Grammar Construction

The workflow to construct Local Grammars is called bootstrapping and consists of alternating refinement and evaluation of the graphs and the concordances they produce. The development of the grammars for this study is based on two main principles. They guide and as well are guided by the author's intuition in order to improve the Local Grammars.

1) For every semantic class of the SHSL_F there is a graph, all put together in the main Local Grammar. Those graphs contain typical context for the respective class as e.g. auxiliary verb constructions or aspect realizations. Typically the development of these graphs begins with a list of verbal expressions that refer to the context as seen in figure 1. All verbs in the big box describe different ways of perceiving phenomena (*seeing* (*ver*), *hearing* (*oir*), but also *understanding* (*endender*), etc.) As seen above, the output of a Local Grammar based on a semantic class not necessarily belongs to that class (e.g. *ejercito dispuesto en formación militar*). This can be ignored if just the extraction of MWE is the goal. If a semantic tagging is desired later, the graphs may need to be refined on this behalf. A manual control is always necessary.
2) The MWE_all graph contains a set of graphs that are meant to reflect all possible syntactical variations of Spanish nominal MWE (Daille, 1994). The set is subdivided into sets for MWE with two, three, four and five main elements. Each subset contains graphs for as much potential syntactical combinations as possible. The example in figure 3 shows the variations for MWE with ANA structure. The upper path connects the three main elements directly or with hyphens (escaped with two #). The bottom path uses quotes at the beginning and the end as additional constraints. The middle path allows the recognition of MWE that begin with roman numerals as *II Guerra Mundial* (World War II). Every local grammar for semantic classes contains links to the MWE_all or several adequate sub graphs which may be addressed separately.

7. Conclusion

I presented a way to extract MWEs using Local Grammars and semantic classes. The approach seems promising especially for religious texts and their literary complexity. This study was applied to the Spanish version of the Hindu poem Bhagavad Gita. The first results based on a very simple Local Grammar are encouraging in terms of quality as well as in terms of quantity. The fact that the basis of text is a complex religious poem does not seem to be a problem for the analysis. Quite to the contrary, the analysis revealed some MWE with religious reference (e.g. *misterio transcendental*, *propósito de la renunciación*), which is interesting for different methods of further text analysis (see below).

Because the Local Grammars and semantical classes used in this study are very limited, the results are limited, too. But based on this data, far better results can be expected for a complete analysis, which needs to include Local Grammars for all semantical classes as well as a refinement of the existing graphs and dictionaries.

8. Future Work

There are several interesting questions arising from the study presented here. The questions are related to the approach itself as well as to the data obtained. Which semantic classes will be the most productive when analysing the Gita? Will those classes change if other (religious or profane) texts are analysed? Which elements need to be included into the Local Grammars to improve the identification of MWE? And although the possibility of semantical tagging of the MWE based on the classes seems to be obvious, is it possible to refine the graphs in a way that a manual control is no longer obligatory?

The religious character of the MWE points to the fact, that MWE can be used for automatic text classification. Is it also possible to assign an analysed text to its according religious community?

As the study is based on a translated text, the following question arises: Are MWE in Spanish translations of MWE in Sanskrit? Does a statistical machine translation system that is going to translate the Gita improve by applying a language model that is trained with MWE? Is the semantic tagging that would be achieved simply by adding the semantic class sufficient to create a special search engine? Also the analysis of the distribution of MWE of other semantical classes as entities or abstracts is interesting (in comparison as well as alone).

9. References

Blanco Escoda, X. (2001). Les dictionnaires électroniques de l'espagnol (DELASs et DELACs). In: *Lingvisticae Investigationes*, 23, 2, pages 201-218.

Blanco Escoda, X. (2010). Etiquetas semánticas de HECHOS como género próximo en la definición lexicográfica. In *Calvo, C. et al. (ed.): Lexicografía en el ámbito hispánico*, pages 159-178.

Breidt, E. et al. (1996). Local Grammars for the Description of Multi-Word-Lexemes and their Automatic Recognition in Texts.

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In: *The Balancing Act, Combining Symbolic and Statistical Approaches to Language -- Proceedings of the Workshop*, pages 29-36.

Gross, M. (1997). The Construction of Local Grammars. In*: Finite-State Language Processing*, pages 329-354.

Guenthner, F. and Blanco Escoda, X. (2004). Multi-lexemic expressions: An Overview. In Leclère, C. et al. (ed.) *Lexique, Syntaxe et Lexique-Grammaire,* pages 239-252.

Paumier, S. (2010): Unitex User Manual 2.1, http://igm.univ-mlv.fr/~unitex/UnitexManual2.1.pdf (Last visit: 02/05/2012)

Shazadi, N. et al. (2011). Semantic Network based Semantic Search of Religious Repository. In: *International Journal of Computer Applications*, 36, 9.

Through Lexicographers' Eyes: Does Morphology Count in Making Qur'anic Bilingual Dictionaries?

Nagwa Younis
Ain Shams University, Egypt
nagwayounis@yahoo.com

Abstract

The existence of several forms of the same word in the Holy Quran is regarded as a source of difficulty for lexicographers who are interested in making Qur'anic bilingual dictionaries. Modern dictionaries nowadays use a considerable amount of corpora to illustrate the actual contexts in which a certain form of a word occurs. The study surveys the attempts of finding equivalents for these forms in the on-line Qur'anic Dictionary provided in the Quranic Corpus (Dukes, 2011). The results of the study shed light on some linguistic aspects in the making of a specialised dictionary of the Holy Quran using a corpus-based approach. These insights are of importance both in the field of lexicography in general and the making of a specialised bilingual Qura'nic dictionary in particular.

Key Words: Corpus-based translation studies, Qur'anic bilingual dictionaries, Corpus Linguistics, Semantic Prosody, Arabic morphology

1. Introduction

Different morphological forms of the same word in Arabic usually convey different meanings that may not be easily rendered into English. The existence of these forms in the Holy Qur'an has been regarded as a source of difficulty for translators who may solve the problem of finding 'equivalent' words for each form by just using the same word more than once. Consider, for example, the difference between نجّى and أنجى in the two verses, (al-A'räf 7:64) and (Yunus 10:73), and how translations did not show the morphological difference between the two forms.

In the above examples, the two English verbs 'delivered' and 'saved' are used interchangeably as equivalents for both Arabic morphological forms أنجى\نجى . Another example is the difference in meaning between the two verbs كسب and اكتسب , where most of the translators rendered the two Arabic forms into the English word 'to earn'. Sharaf and Atwell (2009) maintain that 'the semantic significance of each derivation form is a subtle aspect of Arabic grammar which has no direct equivalent in the grammar/morphology of English or European languages'.

The study sheds light on some attempts of translating these forms in the Holy Qur'an. This is done through scrutinising the Qur'anic Dictionary provided in the Quranic Corpus

(Dukes, 2011). The paper is an attempt to answer the following questions:

1. Does each verb form in the Holy Qur'an convey a different semantic significance, and hence have a different equivalent in English?
2. Is there a specific semantic prosody related to the use of each form in the Holy Qur'an?
3. If so, how can this be accurately detected through corpus analysis? Collocation? (A word is known by the company it keeps?)

2. Research Method

This section is mainly divided into two parts: The first part is a theoretical background in which a brief survey of the major concepts that constitute a backcloth against which different morphological forms of trilateral verbs in the Holy Qur'an are analysed. Such concepts as corpus linguistics and semantic prosody are introduced to pave the way for the second part which is the computational tools used for analysis, i.e. the Quranic Arabic Corpus version 0.4 (Dukes 2011).

2.1 Background

2.1.1 Corpus Linguistics

The word *corpus* was originally used for any collection of writings by a specific author (Baker 1995:225). Nowadays, *corpus* means primarily a collection of texts held in electronic form, capable of being analysed automatically or semi-automatically rather than manually. It includes spoken as well as written texts from a variety of sources, on different topics, by many writers and speakers.

Recently, corpus-based studies have been gaining popularity. Along with the fast developments of corpus linguistics, concordance tools and software also flourish and are constantly being improved. In the areas of translation and lexicography, corpus begins to show its values and advantages. The use of corpus can supplement the deficiencies of conventional methods of translation. Baker (1993: 243) predicted that the availability of large corpora of both original and translated texts, together with the development of a corpus-driven methodology, would enable translation scholars to uncover "the nature of translated text as a mediated communicative event." Since then, a growing number of scholars in translation studies have begun to seriously consider the corpus-based approach as a viable and fruitful perspective within which translation and translating can be studied in a systematic way.

2.1.2. Semantic Prosody

The concept 'semantic prosody' is a rather recent concept that ascribes its parentage to corpus linguistics and its fatherhood to John Sinclair in the University of Birmingham. It refers to the shades of meaning or semantic aura related to the choice of a certain lexical item rather than another (Louw, 1993, Sinclair, 1996, Stubbs, 1996, Partington, 1998, Hunston and Francis, 2000). It reflects the Firthian tradition that lexical items are habitually associated with particular connotations.

Semantic prosody, as I argue elsewhere (Younis 2011), can also be used effectively as a tool for depicting the subtle differences in the meaning of the ‘apparently similar’ structures in the Holy Qur’an. That was applied to the difference in meaning that results from the occurrence of two different prepositions after the same verb in the Holy Qur’an, e.g. ألقى على and ألقى إلى with most translators paying almost no attention to the preposition in favour of the meaning of the main lexical verb. It was suggested in that study that each preposition has a certain semantic prosody that ‘colours’ the meaning of the verb that precedes it. Thus, assigning a certain semantic prosody to a given lexical item can help translators put a clear-cut demarcation line between the possible equivalents of the same item, or otherwise search for additional linguistic tools (e.g. adverbs or adjectives) to illustrate this subtle difference.

2.2 Tools

The study surveys some attempts to give an equivalent to the different morphological representations of the verb forms deriving from the same root in the Holy Qur’an, and how far translators tried to ‘get it right’, as far as these different morphological forms are concerned. This is done through scrutinising the Qur’anic Dictionary provided in the Quranic Corpus version 0.4 (Dukes, 2011). For study limitations, only trilateral verb forms are scrutinised, though nouns and adjectives are very well representative of this phenomenon.

The Quranic Corpus was consulted in this paper on three levels: The first was to identify the different morphological representations of Qur’anic Arabic trilateral verb forms. The second was to search for the different forms of the same stem or root. That was done by selecting a root or base form, and then spotting all different forms deriving from it in the Qur’an with their frequency of occurrence and the context in which they occurred. The third level was analyzing the parallel corpus of the original text with its seven English translations (Sahih International 1997, Pickthall 1930, Yusuf Ali 1934, Shakir 1999, Sarwar 1981, Mohsen Khan 1996 and Arberry 1955). A point that was taken into consideration was the different cultural backgrounds from which those translators came.

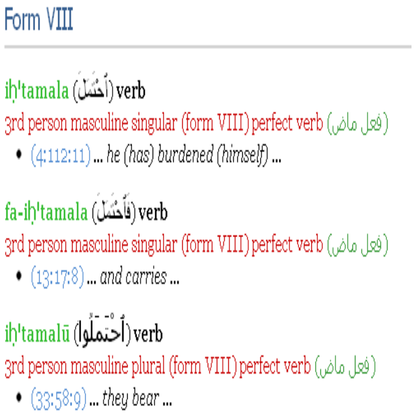
The Quranic Arabic Corpus provides an analysis of the morphological structure of verb forms in the Holy Qur’an based on various treatments of Arabic grammar and morphology (e.g. Ryding 2005, Haywood and Nahmad 1965). A survey was made of all the roots or base forms of trilateral verbs in the Qur’an by means of the Qur’anic Dictionary tool, all the forms that stem from these roots were provided with their translation and frequency of occurrence. Based on this tool, a compiled list of all the verb forms of the same root that had different morphological forms and were rendered into the same English word was provided in the Appendix.

3. Results and Discussion

Various morphological forms of trilateral verbs in the Holy Qur’an fall into two major headings: different forms of the same root and complete versus elided forms of the same verb.

1. Different Forms of the Same Root

In the Holy Qur'an, two or more verbs may share the same root. However they belong to different morphological forms of trilateral verbs as shown in 2.2 above. For example, there was no significant difference shown between the translation of the verb حمل (Form I) and the verb احتمل (Form VIII) as both verb forms were mostly rendered into the English equivalents: *bear* and *carry*.

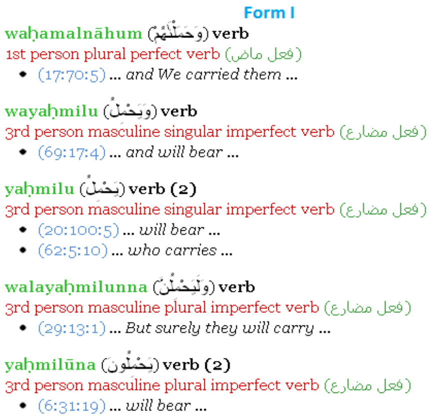


Form VIII

iḥ'tamala (ٱحْتَمَلَ) verb
3rd person masculine singular (form VIII) perfect verb (فعل ماض)
- (4:112:11) ... *he (has) burdened (himself)* ...

fa-iḥ'tamala (فَٱحْتَمَلَ) verb
3rd person masculine singular (form VIII) perfect verb (فعل ماض)
- (13:17:8) ... *and carries* ...

iḥ'tamalū (ٱحْتَمَلُوا۟) verb
3rd person masculine plural (form VIII) perfect verb (فعل ماض)
- (33:58:9) ... *they bear* ...



Form I

waḥamalnāhum (وَحَمَلْنَٰهُمْ) verb
1st person plural perfect verb (فعل ماض)
- (17:70:5) ... *and We carried them* ...

wayaḥmilu (وَيَحْمِلُ) verb
3rd person masculine singular imperfect verb (فعل مضارع)
- (69:17:4) ... *and will bear* ...

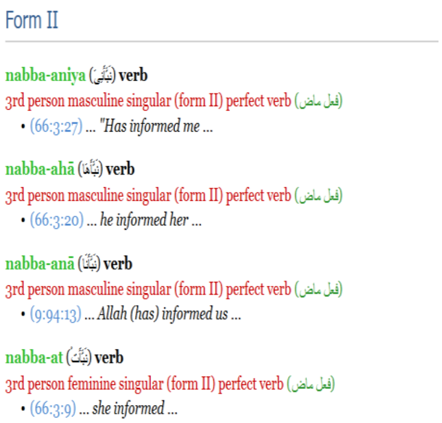
yaḥmilu (يَحْمِلُ) verb (2)
3rd person masculine singular imperfect verb (فعل مضارع)
- (20:100:5) ... *will bear* ...
- (62:5:10) ... *who carries* ...

walayaḥmilunna (وَلَيَحْمِلُنَّ) verb
3rd person masculine plural imperfect verb (فعل مضارع)
- (29:13:1) ... *But surely they will carry* ...

yaḥmilūna (يَحْمِلُونَ) verb (2)
3rd person masculine plural imperfect verb (فعل مضارع)
- (6:31:19) ... *will bear* ...

The semantic prosody of most of the verbs that come under Form VIII (i-F-t-a-3-a-L-a), as opposed to the simple Form I (F-a-3-a-L-a), was mostly related, in the Holy Qur'an, to something done with effort or difficulty. That was clearly the case with اك ت سب (Form VIII) with the preposition ع لى , having the semantic prosody of of 'something hard or done with effort in contradistinction to the verb ك سب (Form I) with the preposition ل , having a positive semantic prosody (cf. Younis 2011).

The same applies to the difference between (Form II) and (Form IV), as in the verbs نبّأ and أنبأ that were both rendered as 'to inform':



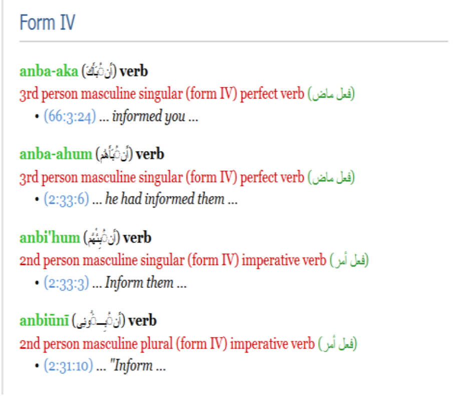
Form II

nabba-aniya (نَبَّأَنِىَ) verb
3rd person masculine singular (form II) perfect verb (فعل ماض)
- (66:3:27) ... *"Has informed me* ...

nabba-ahā (نَبَّأَهَا) verb
3rd person masculine singular (form II) perfect verb (فعل ماض)
- (66:3:20) ... *he informed her* ...

nabba-anā (نَبَّأَنَا) verb
3rd person masculine singular (form II) perfect verb (فعل ماض)
- (9:94:13) ... *Allah (has) informed us* ...

nabba-at (نَبَّأَتْ) verb
3rd person feminine singular (form II) perfect verb (فعل ماض)
- (66:3:9) ... *she informed* ...



Form IV

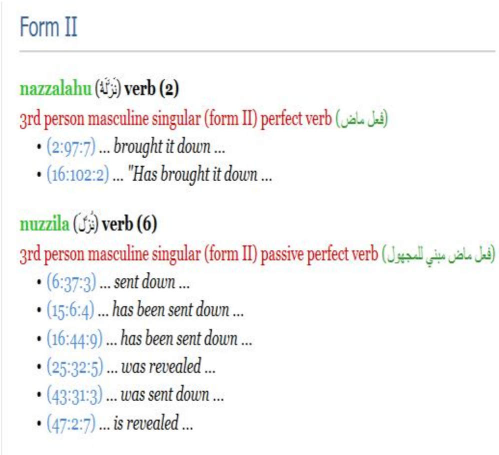
anba-aka (أَنۢبَأَكَ) verb
3rd person masculine singular (form IV) perfect verb (فعل ماض)
- (66:3:24) ... *informed you* ...

anba-ahum (أَنۢبَأَهُم) verb
3rd person masculine singular (form IV) perfect verb (فعل ماض)
- (2:33:6) ... *he had informed them* ...

anbi'hum (أَنۢبِئْهُم) verb
2nd person masculine singular (form IV) imperative verb (فعل أمر)
- (2:33:3) ... *Inform them* ...

anbiūnī (أَنۢبِـُٔونِى) verb
2nd person masculine plural (form IV) imperative verb (فعل أمر)
- (2:31:10) ... *"Inform* ...

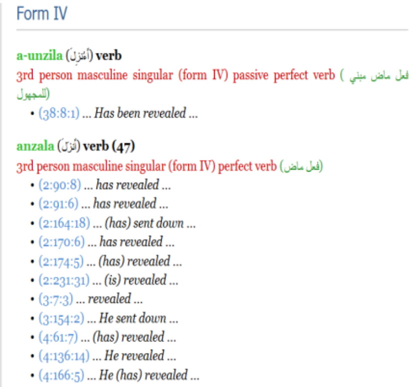
The difference between نزَّل (Form II) and أنزل (Form IV) with two verb forms and two distinctive semantic prosodies does not show in using both ‘revealed’ and ‘sent down’ simultaneously as two equivalents for both forms.



Form II

nazzalahu (نَزَّلَهُ) **verb (2)**
3rd person masculine singular (form II) perfect verb (فعل ماض)
- (2:97:7) ... *brought it down ...*
- (16:102:2) ... *"Has brought it down ...*

nuzzila (نُزِّلَ) **verb (6)**
3rd person masculine singular (form II) passive perfect verb (فعل ماض مبني للمجهول)
- (6:37:3) ... *sent down ...*
- (15:6:4) ... *has been sent down ...*
- (16:44:9) ... *has been sent down ...*
- (25:32:5) ... *was revealed ...*
- (43:31:3) ... *was sent down ...*
- (47:2:7) ... *is revealed ...*



Form IV

a-unzila (أُنزِلَ) **verb**
3rd person masculine singular (form IV) passive perfect verb (فعل ماض مبني للمجهول)
- (38:8:1) ... *Has been revealed ...*

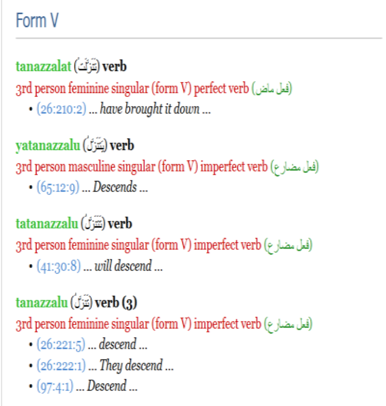
anzala (أَنزَلَ) **verb (47)**
3rd person masculine singular (form IV) perfect verb (فعل ماض)
- (2:90:8) ... *has revealed ...*
- (2:91:6) ... *has revealed ...*
- (2:164:18) ... *(has) sent down ...*
- (2:170:6) ... *has revealed ...*
- (2:174:5) ... *(has) revealed ...*
- (2:231:31) ... *(is) revealed ...*
- (3:7:3) ... *revealed ...*
- (3:154:2) ... *He sent down ...*
- (4:61:7) ... *(has) revealed ...*
- (4:136:14) ... *He revealed ...*
- (4:166:5) ... *He (has) revealed ...*

The form أنزل usually collocates in the Holy Qur’an with the Bible and the Torah that were not sent down gradually over a long period of time. When it occurrs with the Holy Qur’an, it has the same shades of meaning, namely, the revelation of the Holy Qur’an in itself as an action, not the gradual process of revelation that lasted for 23 years.

But the form نزَّل has the habitual co-occurrence with the meaning of ‘gradual process of revelation’, as appears in (3:3).

2. Complete Vs. Elided Forms of the Same Verb

The same morphological base forms of trilateral verbs have differential realizations in the Holy Qur’an (e.g. Form I, II, III,…). Nevertheless, some verbs appear in their complete forms, others elided, i.e. one of their letters is omitted. For example, the verb تنزل occurs in the Holy Qur’an six times in Form V (t-a-F-a-33-a-L-a). Only three of them were in their complete form; the other three were elided. تتنزل is in the present tense, and is inflected for feminine gender . تنزل , on the other hand, is the elided form with the double تاء becoming only one.



Form V

tanazzalat (تَنَزَّلَتْ) **verb**
3rd person feminine singular (form V) perfect verb (فعل ماض)
- (26:210:2) ... *have brought it down ...*

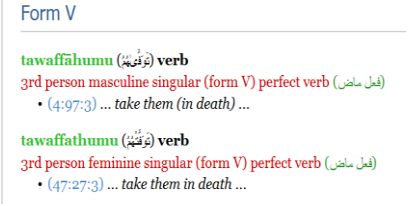
yatanazzalu (يَتَنَزَّلُ) **verb**
3rd person masculine singular (form V) imperfect verb (فعل مضارع)
- (65:12:9) ... *Descends ...*

tatanazzalu (تَتَنَزَّلُ) **verb**
3rd person feminine singular (form V) imperfect verb (فعل مضارع)
- (41:30:8) ... *will descend ...*

tanazzalu (تَنَزَّلُ) **verb (3)**
3rd person feminine singular (form V) imperfect verb (فعل مضارع)
- (26:221:5) ... *descend ...*
- (26:222:1) ... *They descend ...*
- (97:4:1) ... *Descend ...*

It should be noted that the frequency of occurrence of the reduced form تنزَّل usually accompanies short events or the events that took almost no time. There seemed to be a kind of correspondence between the time duration of the event and the time duration of the pronunciation of the verb denoting the event (cf. al-Samarra'i 2006). For example, in (97:4) the elided verb form تنزّل refers to a short event that happened only once a year; in the night of decree, whereas in (41:30) the complete verb form تتنزّل refers to an event that is possibly recurrent at every moment. Since the phenomenon of eliding verb forms does not have an equivalent in English, the translations did not use any linguistic tool to highlight the semantic difference between the complete form and the elided one.

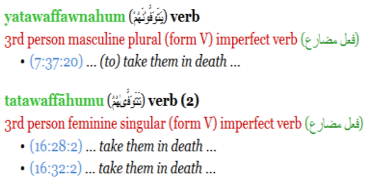
Another example is the verb تتوفاهم (Form V) which also occurs in the elided form توفاهم with the omission of تاء at the beginning of the verb.



Form V

tawaffāhumu (تَوَفَّىٰهُمُ) **verb**
3rd person masculine singular (form V) perfect verb (فعل ماض)
- (4:97:3) ... *take them (in death)* ...

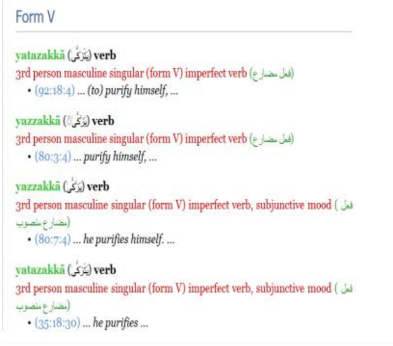
tawaffathumu (تَوَفَّتْهُمُ) **verb**
3rd person feminine singular (form V) perfect verb (فعل ماض)
- (47:27:3) ... *take them in death* ...



yatawaffawnahum (يَتَوَفَّوْنَهُمْ) **verb**
3rd person masculine plural (form V) imperfect verb (فعل مضارع)
- (7:37:20) ... *(to) take them in death* ...

tatawaffāhumu (تَتَوَفَّىٰهُمُ) **verb (2)**
3rd person feminine singular (form V) imperfect verb (فعل مضارع)
- (16:28:2) ... *take them in death* ...
- (16:32:2) ... *take them in death* ...

The verb يتزكى also occurs in the elided form يزَّكى with the omission of the تاء . Both forms are translated 'purify', without any distinction whatsoever between the complete and the elided forms.



Form V

yatazakkā (يَتَزَكَّىٰ) **verb**
3rd person masculine singular (form V) imperfect verb (فعل مضارع)
- (92:18:4) ... *(to) purify himself,* ...

yazzakkā (يَزَّكَّىٰ) **verb**
3rd person masculine singular (form V) imperfect verb (فعل مضارع)
- (80:3:4) ... *purify himself,* ...

yazzakkā (يَزَّكَّىٰ) **verb**
3rd person masculine singular (form V) imperfect verb, subjunctive mood (فعل مضارع منصوب)
- (80:7:4) ... *he purifies himself.* ...

yatazakkā (يَتَزَكَّىٰ) **verb**
3rd person masculine singular (form V) imperfect verb, subjunctive mood (فعل مضارع منصوب)
- (35:18:30) ... *he purifies* ...

Looking deeply into the attempts of various translations of the two versions of the verb تتزكى with its complete and elided forms, it was observed that no additional linguistic tool was used to differentiate between them.

It is worth-mentioning that despite all the great efforts exerted in the Qur'anic Corpus (2011), a corpus-based linguistic investigation should be conducted very carefully. For example, the verb يفترين in (12:23) was classified as Form I (Fa3aLa), when in fact it belongs to Form VIII (i-F-t-a-3-a-L-a).

4. Conclusion

Scruitinsing the translations provided for the different realizations of verb forms in the Qur'anic Corpus (Dukes, 2011), it was observed that these realizations were almost translated the same. The argument postulated in this study is that corpus research can provide two layers of solution for the translation of different morphological forms in the Holy Qur'an: The first layer is the semantic significance associated with the realization(s) of the verb form itself. It was suggested in the theoretical part that each of the twelve morphological forms of trilateral verbs in the Holy Qur'an has its own connotation(s). The second layer is the semantic prosody related to the use or contextual use of each form. This typically applies to both categories, i.e. verb forms that have the same root and complete versus elided forms of the same verb.

This study is an attempt to suggest that corpus research is of importance to translators as it gives more insights into and a wider scope of rendering Arabic, in general, and Qur'an, in particular, into other languages. Corpus here mainly opens a window to see more clearly all linguistic problems with a view to finding more objective solutions.

References

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, Cx Francis&.E. Tognini-Bonelli (eds.), *Text and Technology: In Honor of John Sinclair*. Amsterdam & Philadelphia: John Benjamins, pp. 233-250.

Baker, M. (1995). *Corpora in translation studies: an overview and some suggestion for future research*. Target 7(2), pp. 223-43.

Dukes, K. (2011). *Quranic Corpus*. School of Computing, University of Leeds. http://corpus.quran.com/

Firth, J. (1968). A Synopsis of Linguistic Theory 1930-1955. In Palmer, (Ed.) *Selected Papers of J.R. Firth.* Harlow: Longman.

Haywood, J. & Nahmad, H. (1965). A New Arabic Grammar of the Written Language. second edition. London: Lund Humphries.

Hunston, S, and G. Francis. 2000. *Pattern grammar. A corpus-driven approach to the lexical grammar of English.* Amsterdam and Philadelphia: John Benjamins.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair*. Amsterdam: John Benjamins, pp. 157-176.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. Philadelphia, PA: John Benjamins.

Samarra'i, F. (2006). Balaghatu-lkalema fi-tta'beer al-qur'aani. *The Rhethoric of the Word in the Qur'anic Expression*. 2nd edition. Cairo: Atek Publishing Company.

Sharaf, A. and Atwell, E. (2009). A Corpus-based computational model for knowledge representation of the Qur'an. *5th Corpus Linguistics Conference*, Liverpool.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Sinclair, J. (1996). The search for units of meaning. *TEXTUS: English Studies in Italy, 9*(1), pp.75-106.

Sinclair, J. (2004). *Trust the Text: Language, Corpus and Discourse.* London: Routledge.

Stubbs, M. (1996). *Text and Corpus Linguistics*. Oxford: Blackwell.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Setting*. Glasgow: CUP.

Ryding, K. (2005). *A Reference Grammar of Modern Standard Arabic*. Cambridge: CUP.

Younis, N . (2011). *Semantic Prosody as a Tool for Translating Prepositions in the Holy Qur'an.* Workshop on Arabic Linguistics. Lancaster University. http://ucrel.lancs.ac.uk/wacl/abstracts.pdf

Appendix

Frequency of Occurrence of Some Trilateral Verbs in the Holy Qur'an that Share the Same Root but with Different Morphological Forms, and are translated the same in the Quranic Dictionary (Dukes 2011)

Verb	Root	Form	Frequency	Translation
بلّغ	ب ل غ	II	6	to convey
أبلغ	ب ل غ	IV	5	to convey
بلو	ب ل و	I	20	to test
يبلى	ب ل و	IV	2	to test
ابتلى	ب ل و	VIII	8	to test, to try
بوأ	ب و ا	II	6	to settle
يتبوأ	ب وا	V	4	to settle
تبع	ت ب ع	I	9	to follow
أتبع	ت ب ع	IV	15	to follow
اتّبع	ت ب ع	VIII	136	to follow
ثقُلت	ث ق ل	I	4	to be heavy
أثقلت	ث ق ل	IV	1	to be heavy
اثاقل	ث ق ل	VI	1	to be heavy
ثوّب	ث و ب	II	1	to reward
أثاب	ث و ب	IV	3	to reward
جرَح	ج ر ح	I	1	to commit
اجترح	ج ر ح	VIII	1	to commit

Hybrid Approach for Extracting Collocations from Arabic Quran Texts

Soyara Zaidi[1], Ahmed Abdelali[2], Fatiha Sadat[3], Mohamed-Tayeb Laskri[1]

[1]University of Annaba, Algeria
[2]New Mexico State University, USA
[3]UQAM, Canada
E-mail: [1](soraya.zaidi, laskri)@univ-annaba.org, [2]aabdelali@research.nmsu.edu, [3]sadat.fatiha@uqam.ca

Abstract

For the objective of building Quran Ontology using its original script, existing tools to develop such resource exists to support languages such as English or French hence can't be used for Arabic. In most cases, significant modifications must be made to obtain acceptable results. We present in this paper an automatic approach to extract simple terms and collocations to be used for the ontology. For extracting collocations, we use a hybrid approach that combines linguistic rule-based method, and Mutual-Information-based approach. We use a mutual information-based approach to filter, enhance and further improve the precision of the results obtained by linguistic method. Extracted collocations are considered essential domain terms to build Arabic ontology of Quran. We use The Crescent Quranic tagged Corpus which consisted of Quran text tagged per word, verse and chapter; it contains as well additional information about morphology and POS and syntax.

Keywords: Collocations Extraction, Linguistic approach, Mutual Information, Ontology Construction, Quran

1. Introduction

Quran, Islam's holy book, was revealed to the Prophet Muhammad (PBUH) in Arabic; a language that is a member of the family of languages called Semitic, "*a term coined in the late eighteenth century CE and inspired by the Book of Genesis's account of the physical dispersal of the descendants of Noah together with the tongues they spoke*"(Shah, 2008). This group of languages shares a number of features including triliterality (majority of Arabic words are based on triliteral root), the appendage of morphological markers and inflection; where the general meaning and category of the morpheme tend to be preserved; while grammatical information could change. Great significance was attached to Arabic language as the most dynamic surviving Semitic languages while preserving a wealth of linguistic information connected to its early development and history. The revelation of Quran in Arabic; gave the language a great religious significance and was the main motive that carried this language beyond its native geographic space. Arabic has 28 letters and is written from right to left. Arabic language processing is considered complex because of structural and morphological characteristics, such as derivation, inflection, and polysemy. As much as these characteristics are features and distinctive treats for Arabic, they do present a challenge for automatic processing of the language namely named entity extraction.

Quran contains 114 surah (chapters), 6346 verses, 77 439 words and 321 250 characters. As a divine word of God, Quran is very central to the religion of Islam. The holy book contains guidance, jurisprudence, morals, and rituals; in addition to some accounts of historical events. Since the early dates, scholars of Islam produced a large number of commentary and explication (tafsir), for the purpose of explaining the meanings of the Quranic verses, clarifying their importance and highlighting their significance. With the advances in technology, searching religious text and resources is being made easier and accessible via new mediums. Attempts were made manually compile Quranic Ontology for the purpose of representing knowledge "in traditional sources of Quranic analysis, including the tradition of the prophet muhammad, and the tafsīr (Quranic exegesis) of ibn kathīr". These efforts are cumbersome and require a lot of efforts and time. Dukes et al. (2010) **Error! Reference source not found.**compiled a network of 300 linked concepts with 350 relations. See Figure 1. The first step in automating this process is to extract terms and collocations that will form the concepts in the ontology (Roche et al., 2004; Schaffar, 2008; Seo & Choi, 2007).

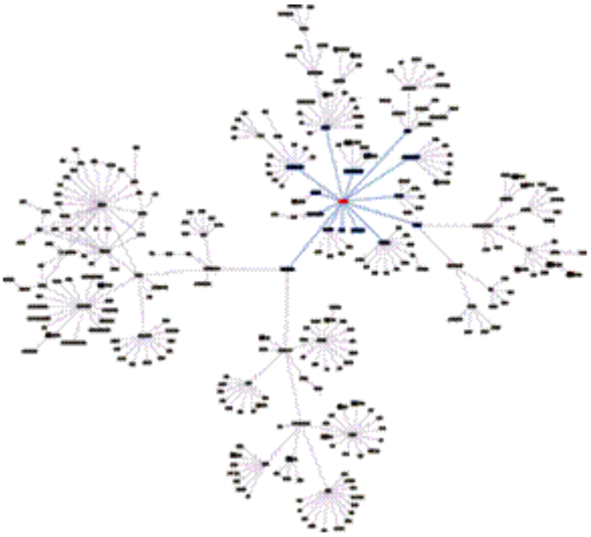


Figure 1: Concepts and Relations in the Quranic Ontology. Reprinted from http://corpus.quran.com

Collecting polylogs, known units or what is referred to by collocations were initiated by Language Teachers who noted their pedagogical value and usefulness in Language Teaching. Early work on collocations were documented in dictionaries such as Idiomatic and Syntactic English Dictionary (Hornby, 1942), The Oxford Advanced Learner's Dictionary (Hornby et al, 1948) and The Advanced Learner Dictionary of Current

English (Hornby et al., 1952). Even though collocations were seen as pure statistical phenomenon of word co-occurrence; it is well accepted that they are more a syntactical-bound combination (Seretan, 2005). A number of languages have been investigated using linguistic methods or statistical techniques which showed interesting results such as the log-likelihood ratio measure (LLR) (Shah, 2008). LLR method was widely accepted and used by numerous researchers as one of the most useful techniques thanks to its particular suitability for low-frequency data.

Due to the linguistic feature of every language, tools developed are mostly language oriented. A number of tools were proposed to extract terms in languages such as French, English or German. Some of these tools can easily be adapted with some minor modifications for other language such as Arabic language namely statistical or hybrid tools. Linguistic tools may require moderate to extensive work to be successfully adopted.

The remainder of the paper is organized as follows; the second section of this paper, we present a statistical approach for Arabic simple terms extraction. Section 3 introduces a linguistic approach for extracting Arabic collocations. In Section 4, we discuss the filtering result with Mutual Information-Based approach. We provide an assessment and evaluation in term of Precision, Recall and F.measure, in section 5; and finally, the conclusion details future and some perspectives for our work.

2. Statistical Approach for Extracting Simple Terms from Quranic Corpus

2.1 Extracting Arabic Terms

Terms extraction is the first step in the construction of ontology (Roche, et al., 2004; Schaffar, 2008); these terms represent the semantic concepts that forms the ontology. The types of terms can be either simple or compound. We will present a statistical approach for the extraction of simple terms. The simple terms are difficult to recognize because they carry less meaning than compound terms, however we cannot bypass them for their importance and they are a key for the representation of the domain. To extract these terms we chose an approach based on tf-idf measurement. The idea behind the approach to use the frequency measured in the form of tf-idf is inferred from Information Theory based research and more precisely Zipf law (Roche & Kodratoff, 2006; Zipf, 1949) which concluded that the terms the most informative of a corpus are not the most frequent words; although such words are mostly useful words; in the other hand the less frequent words are not the most informative words neither. These words can be spelling errors or words very specific to a document in the corpus.

2.2 Design

After morphologically analyzing the text with Aramorph[1], the result of this analysis is used to calculate the frequency of each word in all chapters of the Quranic corpus where the terms appear. After the weights were computed using tf-idf "term frequency - inverse document frequency" formula, each word has its weight assigned. The terms are listed in descending order of weighting and threshold is set experimentally to determine the list of words to retain or reject. We should note that we used documents from different sources than Quran in order to eliminate the frequent words. These words occur in various categories and they don't carry a distinctive meaning; such terms will not be retained. The additional documents of different categories are selected from Al-Sulaiti corpus, which consists of over 843,000 words in 416 files covering a wide range of categories (Al-Sulaiti & Atwell, 2006) The usage of the additional document was necessary to supplement the short chapter in Quran. Where these chapters –Suras- contain few sentences and the measurement of the words will almost look alike.

The measurement tf-idf is expressed as the following:

For a given term i in a document j among N documents of the collection or the corpus (Béchet, 2009).

$$w_{ij} = tf_{ij} \times idf_i$$

such that:

$$idf_i = log\frac{N}{n_i}$$

Where w_{ij} is the weight of the term i in the document j;tf_{ij} is the frequency of the term i in the document j; N is the total number of documents in the corpus; n_i is the number of documents where the term i appears. Using log nullify the effects of the terms that appears in every document which is an indication that the terms is not related to a specific domain. Figure 2 summarizes the process and the tools used for term extraction.

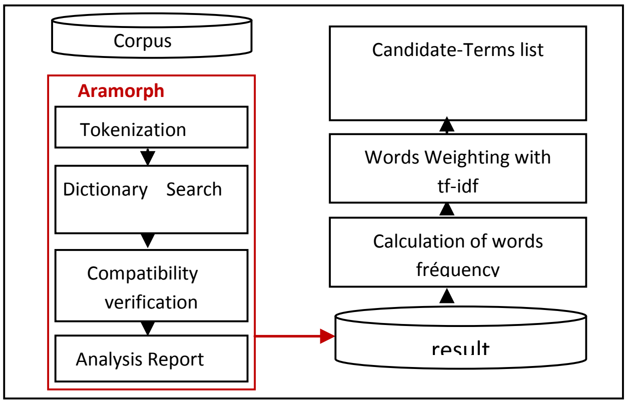
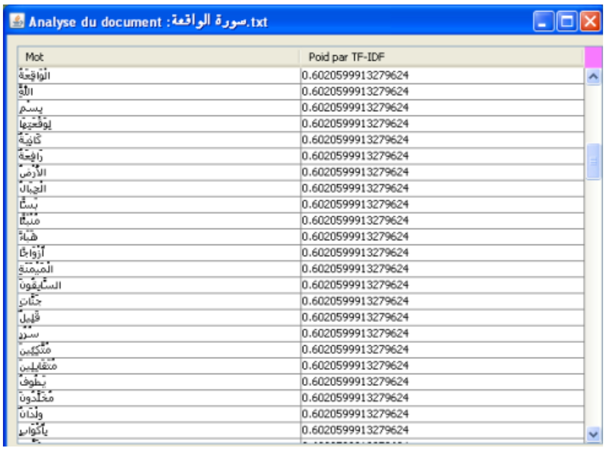


Figure 2: Simple terms extraction

Once the weights get computed, we proceed to sort the words in a decreasing order; a threshold is fixed to eliminate the words with a frequency that didn't surpass the threshold. A final list is presented to the user or an expert to validate the selected words – See Figure 3.

[1] www.nongnu.org/aramorph/french



Analyse du document :سورة الواقعة.txt

Mot	Poid par TF-IDF
الْوَاقِعَةُ	0.6020599913279624
اللَّهِ	0.6020599913279624
بِسْمِ	0.6020599913279624
لِوَقْعَتِهَا	0.6020599913279624
كَاذِبَةٌ	0.6020599913279624
رَافِعَةٌ	0.6020599913279624
الْأَرْضُ	0.6020599913279624
الْجِبَالُ	0.6020599913279624
بَسًّا	0.6020599913279624
مُنْبَثًّا	0.6020599913279624
هَبَاءً	0.6020599913279624
أَزْوَاجًا	0.6020599913279624
الْمَيْمَنَةِ	0.6020599913279624
السَّابِقُونَ	0.6020599913279624
جَنَّاتِ	0.6020599913279624
قَلِيلٌ	0.6020599913279624
سُرُرٍ	0.6020599913279624
مُتَّكِئِينَ	0.6020599913279624
مُتَقَابِلِينَ	0.6020599913279624
يَطُوفُ	0.6020599913279624
مُخَلَّدُونَ	0.6020599913279624
وِلْدَانٌ	0.6020599913279624
بِأَكْوَابٍ	0.6020599913279624

Figure 3: Extracted terms with their tf-idf weight

2.3 Evaluation

Khadhr (2005) manually compiled a list of 17622 words published in his book "*Lexicon of Quran Words*". The list was used as a reference to measure the performance of our approach; we obtained the following results summarized in Table 1.

Precision	Recall	F.Measure
0.88	0.92	0.89

Table 1: Recall/Precision results

From the results listed in Figure 3 we notice that words such as "أول" (*first*), "نحن (*we*) rejected and "كأمثال "(*as examples*) should not be included as terms in the analysis phase, and they were not removed because they were considered as noun with an important weight. That was reflected by the Precision at 88% Therefore, the Noise seems to be considerable (11%) which provides an opportunity for improvement for the approach.

3. Linguistic approach for extracting Arabic collocations from Quranic corpus

3.1 JAPE Language

JAPE (Java Annotation Pattern Engine) provides finite state transduction over annotations based on regular expressions. JAPE is a version of CPSL (Common Pattern Specification Language) (Maynard at al., 2002). A JAPE grammar consists of a set of phases, each of which consists of a set of pattern/action rules. The phases run sequentially and constitute a cascade of finite state transducers over annotations. The left-hand-side (LHS) of the rules consists of an annotation pattern description. The right-hand-side (RHS) consists of annotation manipulation statements. Annotations matched on the LHS of a rule may be referred to on the RHS by means of labels that are attached to pattern elements (Thakker et al., 2009).

3.2 Adapting Gate to Extract Collocations

Collocations include noun phrases like *hard task* and *balanced ecosystem* or *hard-earned money*, phrasal verbs like to *be in touch* or to *make progress*, and other phrases like *state-of-the-art.*

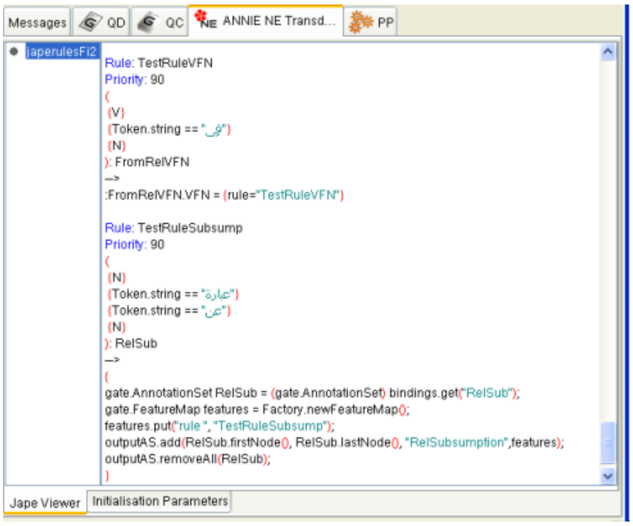
Gate**Error! Reference source not found.**was built mainly for named entities extraction; our idea is to adapt it for collocations extraction from Quranic tagged corpus using predefined patterns in similar fashion as experiments of Zaidi et al. (2010). Extracting terms automatically from corpora is to identify elements carrying a key concept in the specified domain. Terms may be simple or compound words form as Noun-Noun: “تجارة_ حاضرة” (*present_trade*), Verb-Noun " اتقوا_الله» (*fear_God*), Noun-Adjective:” رهان_مقبوضة” (*bet_receipt*), Noun-preposition-Noun: " شعوبا_و_قبائل " (*nations and tribes*).

After choosing the pattern, for example, Noun-Adjective, we write the appropriate Jape rule, then create a new NE ANNIE transducer, the rule is passed as parameter to the transducer, collocations in Arabic text will be detected, displayed and the annotated document can be saved in datastore with the new tag or label in XML format which could be exploited for other tasks or used as an input for another experiment.

We have experimented with the transducer on the both corpora to get a list of domain terms. Using Jape rules we define new annotations used to detect collocations such as Noun-Adjective Noun-Noun, Verb-Noun …etc.

Each Jape rule grammar has one of 5 possible control styles: ‘brill’, ‘all’, ‘first’, ‘once’ and ‘appelt’. This is specified at the beginning of the grammar (Figure 4).

The Brill style means that when more than one rule matches the same region of the document, they are all fired. The result of this is that a segment of text could be allocated more than one entity type, and that no priority ordering is necessary.



```
Rule: TestRuleVFN
Priority: 90
(
 {V}
 {Token.string == "في"}
 {N}
): FromRelVFN
-->
:FromRelVFN.VFN = {rule="TestRuleVFN"}

Rule: TestRuleSubsump
Priority: 90
(
 {N}
 {Token.string == "عبارة"}
 {Token.string == "عن"}
 {N}
): RelSub
-->
{
gate.AnnotationSet RelSub = (gate.AnnotationSet) bindings.get("RelSub");
gate.FeatureMap features = Factory.newFeatureMap();
features.put("rule ", "TestRuleSubsump");
outputAS.add(RelSub.firstNode(), RelSub.lastNode(), "RelSubsumption",features);
outputAS.removeAll(RelSub);
}
```

Figure 4: Simple Jape rule for detecting Verb-Noun collocations

The ‘all’ style is similar to Brill, in that it will also execute all matching rules, but the matching will continue from the next offset to the current one. The ‘first’ style, a rule fires for the first match that’s found. This makes it inappropriate for rules that end in ‘+’ or ‘?’ or ‘*’. Once a match is found the rule is fired; it does not attempt to get a longer match (as the other two styles do). In the style ‘once’, when a rule has fired, the whole

JAPE phase exits after the first match.
Where in the appelt style, only one rule can be fired for the same region of text, according to a set of priority rules. Priority operates in the following way.

1. From all the rules that match a region of the document starting at some point X, the one which matches the longest region is fired.
2. If more than one rule matches the same region, the one with the highest priority is fired.
3. If there is more than one rule with the same priority, the one defined earlier in the grammar is fired (Maynard at al., 2002).

The previous rule in Figure 3 allows recognizing words in a text with a Noun tag followed by adjective tag, to give out the collocation consisting of N-ADJ, similarly we can write additional rules to recognize other types collocations. On the LHS, each pattern is enclosed in a set of round brackets and has a unique label; on the RHS, each label is associated with an action. In this example, the annotation is labelled "SomeLabel" and is given the new annotation N_ADJ.
The creation of new transducers with the previous settings, will allow identifying collocations according to specified syntactic patterns (See Figure 5).
A validation by human expert in the domain is carried after. This is consists of accepting or rejecting collocations displayed, because it is possible to get words that validate the pattern but the obtained set is not considered as a relevant collocation.

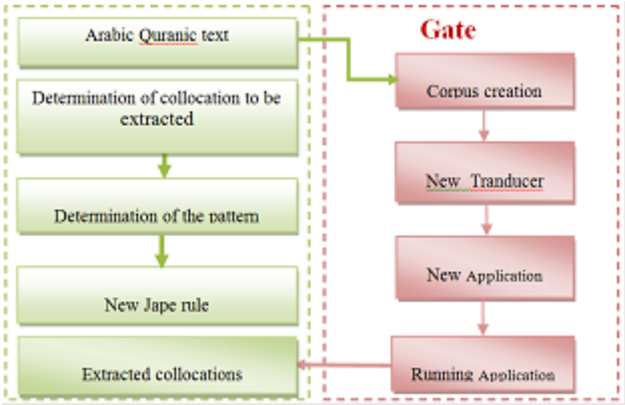
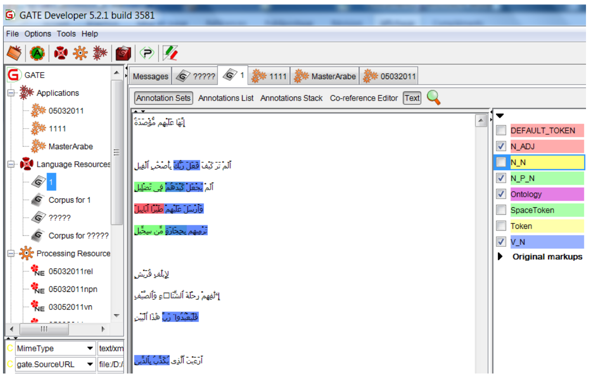


Figure 5: Steps to extract Arabic Quranic collocation with GATE

3.3 Results and Discussions

The aim of our work is to extract collocations from Quranic text. After validation, a list of relevant collocations-these are considered as terms of the domain-will be selected in order to use them in automatic building of Quran ontologie. After running application with new parameters we can display extracted collocations as shown below (Figure 6). The collocations extracted as "العهن المنفوش" (*fluffed wool*), "الفراش المبثوث" (*scattered moth*), "عيشة راضية" (*pleasant life*) can be displayed or saved in datastore for future use. A vital part of any language engineering application is the evaluation of its performance, in order to evaluate the system performance; we used traditional measures of Precision and Recall (and then F.measure) on our corpus.



☐ DEFAULT_TOKEN
☑ N_ADJ
☐ N_N
☑ N_P_N
☑ Ontology
☐ SpaceToken
☐ Token
☑ V_N

إِنَّهَا عَلَيْهِم مُّؤْصَدَةٌ

أَلَمْ تَرَ كَيْفَ فَعَلَ رَبُّكَ بِأَصْحَٰبِ ٱلْفِيلِ

أَلَمْ يَجْعَلْ كَيْدَهُمْ فِى تَضْلِيلٍ

وَأَرْسَلَ عَلَيْهِمْ طَيْرًا أَبَابِيلَ

تَرْمِيهِم بِحِجَارَةٍ مِّن سِجِّيلٍ

لِإِيلَٰفِ قُرَيْشٍ

إِۦلَٰفِهِمْ رِحْلَةَ ٱلشِّتَآ□ءِ وَٱلصَّيْفِ

فَلْيَعْبُدُوا۟ رَبَّ هَٰذَا ٱلْبَيْتِ

أَرَءَيْتَ ٱلَّذِى يُكَذِّبُ بِٱلدِّينِ

Figure 6: Quranic collocations extracted using JAPE rules

For this step, We selected a representative number of chapters (Long, Meduim and Short Suras) and were manually annotated by domain expert. Similarly, we run the same chapters though GATE using the described Jape rules. AnnotationDiff tool (Dukes at al., 2010; Maynard & Aswani, 2009) enables to compare two sets of annotations on a document, in order to compare a system-annotated text with a reference (hand-annotated) text, We annotated manually the same documents which are annotated using Jape rules with Gate, and then we used AnnotationDiff to calculate Precision, Recall then F.Measure.

Precision	Recall	F.Measure
0.50	1.00	0.66

Table 2: Recall/Precision Result

We know that the Recall is inversely proportional to Precision, then if we want to improve the first we reduce the second and vice-versa, therefore F.measure is better to give an idea about the performance of chosen method. F.measure gives 66%, although may seem comparable to peer results such in (Maynard at al., 2002). This is mainly due to the ambiguity, example: الجاهل أغنياء (*the ignorant self-sufficient*) in the verse "يَحْسَبُهُمُ الْجَاهِلُ أَغْنِيَاءَ مِنَ التَّعَفُّفِ" (*think about them the ignorant self-sufficient because of restraint*) where أغنياء (*self-sufficient*) is tagged as adjective for the noun الجاهل (*the ignorant*) usually أغنياء (*self-sufficient*) is an adjective, but in this verse, it is used as noun, this is because, in Arabic there are transitive verbs which need more than one object noun and here the first object noun is not mentioned in the regular expected order. Another case is "و أطعنا غفرانك" (*and we obeyed, forgiveness*) annotated as Verb-Noun, where غفرانك (*forgive us*) is the first word in the next sentence غفرانك ربنا (*forgiveness our lord*)and doesn't belong to the same sentence. It is important to notice that Quran corpus is quite unique and special in its style, characteristics and features, co-occurrence of two words even though common does not mean necessarily that they form a relevant collocation. Finally, tokens in the used corpus are tagged either as noun, verb or preposition; more detailed tagging will certainly improve precision.

4. Filtering Result with Mutual Information-Based Approach

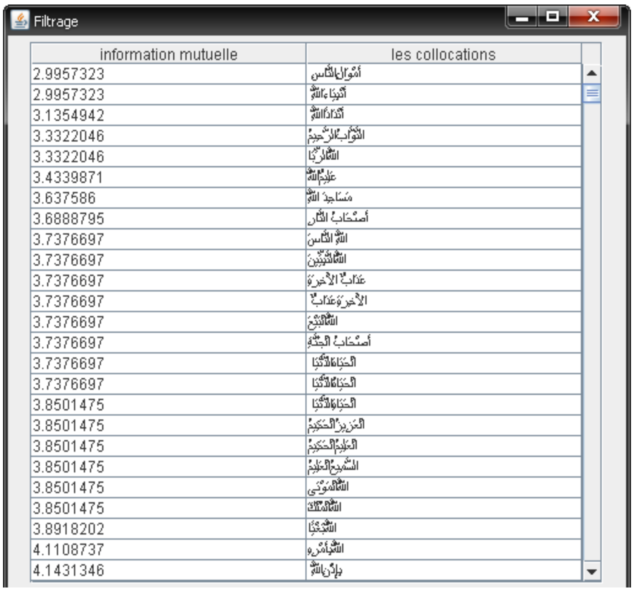
To further improve precision, we tried to filter results with a statistical approach based on Mutual Information (MI). Mutual information is one of many quantities that measures how much one random variable tells us about another. High mutual information indicates a large reduction in uncertainty; low mutual information indicates a small reduction; and zero mutual information between two random variables means the variables are independent (Latham & Roudi, 2009).
For two variables x and y whose joint probability distribution is Pxy(x,y), e mutual information between them, denoted I(x, y), is given by (Bernard, 2006; Church & Hanks, 1990):

$$I\,(x,\,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Where

$$P(x,\,y) = \frac{f(x,y)}{N};\ P(x) = \frac{f(x)}{N};\ P(y) = \frac{f(y)}{N}$$

P (x, y), P (x) and P (y) are estimated by maximum likelihood where f (x) is the frequency of the word x in a corpus of size N (number of occurrences of words). According to the formula, if x and y are related, P (x, y) is greater than P (x). P (y), with the hypothesis I(x,y)=0 if x and y are independent, I (x, y) ≥ 0.
After calculating I(x,y) for each couple (x,y)extracted with the previous approach, we set a threshold experimentally, If the result ≥ threshold, x and y are considered as a compound term else the couple is rejected (See Figure 7).



information mutuelle	les collocations
2.9957323	أَمْوَالَالنَّاسِ
2.9957323	أَنبِيَاءَاللَّهِ
3.1354942	أَندَادًااللَّهِ
3.3322046	التَّوَّابُالرَّحِيمُ
3.3322046	اللَّهَرَبَّنَا
3.4339871	عَلِيمٌاللَّهِ
3.637586	مَسَاجِدَ اللَّهِ
3.6888795	أَصْحَابُ النَّارِ
3.7376697	اللَّهِ النَّاسَ
3.7376697	اللَّهَالْمُتَّقِينَ
3.7376697	عَذَابُ الْآخِرَةِ
3.7376697	الْآخِرَةِعَذَابٌ
3.7376697	اللَّهَالْبَيْعَ
3.7376697	أَصْحَابُ الْجَنَّةِ
3.7376697	الْحَيَاةَالدُّنْيَا
3.7376697	الْحَيَاةُالدُّنْيَا
3.8501475	الْحَيَاةِالدُّنْيَا
3.8501475	الْعَزِيزُالْحَكِيمُ
3.8501475	الْعَلِيمُالْحَكِيمُ
3.8501475	السَّمِيعُالْعَلِيمُ
3.8501475	اللَّهُالْمَوْتَى
3.8501475	اللَّهُالْمُلْكَ
3.8918202	اللَّهُشَيْئًا
4.1108737	اللَّهُبِأَمْرِهِ
4.1431346	بِإِذْنِاللَّهِ

Figure 7: Collocations with their Mutual information

Then, we calculate the precision. The results presented in Table 3.

Approach	Precision
linguistic approach (before filtering)	0.50
statistical approach (after filtering)	0.86

Table 3: Precision after filtering with statistical method

The obtained results shows a clear improvement when compared to the previous experiment –pure linguistic, no filtering-. This means that filtering with the statistical method has a clear impact on the accuracy and the choice we made by combining the two approaches is more than justified. If needed, the result can be validated and further improved by domain human expert to get more refined list of collocations.

5. Conclusion and Perspectives

We have presented in this paper the problem of extracting collocations from Quranic texts. First we tested a statistical approach which produced very acceptable results. The biggest challenge was the short suras where the size is too small to have distinctive measurement between the words. Using the additional corpus has helped; but the results could be improved further using comparable corpus in contrast to the contemporary Arabic used. We used as well a linguistic approach to extract named entities, which produced lower results .We additionally used a Mutual Information-based approach to improve the results obtained by the linguistic approach. This step has further improved the precision from 0.5 to 0.86. , the quality and the size of the corpus was among the major challenges, we are currently working to improve the tagging and trying to obtain a larger tagged corpus such as the Penn Arabic Treebank or comparable corpora to validate our approach and allow extracting more terms and collocations. On the other hand, work is still ongoing, on relation extraction using syntactic patterns; by building new patterns and templates and refining existing ones. Our overall objective is to build a semantic network with extracted terms and relations that could replace the manually crafted and limited existing ontologies. The resulted ontology will be a platform to improve information retrieval in Quranic text and it can be used to improve machine translation and automatic indexing as well.

6. References

Al-Sulaiti, L., Atwell, E. (2006). The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, vol. 11, pp. 135-171..

Béchet N., (2009). Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes, l'Université de sciences et Techniques du Languedoc, thèse pour obtenir le diplôme de doctorat.

Bernard , D. (2006). Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales, Thèse pour obtenir le grade de docteur, université Joseph Fourier Grenoble1.

Church, K. W., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 16(1):22–29.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: An Architecture For Development of Robust HLT Applications. Proceedings of The 40th Annual Meeting of The Association For Computational Linguistics (ACL), (Pp. 168-175). Philadelphia.

Dukes, K., Atwell, E., & Sharaf, A. (2010). Syntactic Annotation Guidelines For The Quranic Arabic Treebank. The Seventh International Conference on Language Resources and Evaluation (LREC-2010). Valletta, Malta.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19.1 (Mar. 1993), 61-74.

Hornby, A. S. (1942). Idiomatic and syntactic English dictionary. Tokyo: Institute for Research in Language Teaching.

Hornby, A. S., Cowie, A. P., & Lewis, J. W. (1948). Oxford advanced learner's dictionary of current english. London: Oxford University Press.

Hornby, A. S., Gatenby, E. V., & Wakefield, H. (1952). The Advanced learner's dictionary of current English. London: Oxford University Press.

Khadhr, Z. M. (2005) Qictionary of Quran Words. Retrieved from http://www.al-mishkat.com/words/

Latham, P., & Roudi, Y. (2009), Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't. PLoS Comput Biol. 2009 May; 5(5).

Maynard, D., & Aswani, N. (2009). Annotation and Evaluation. Tutorial Summer School Sheffield.

Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., & Wilks, Y. (2002). Architectural elements of language engineering robustness. Natural Language Engineering, 8(2): 257-274. Cambridge University Press. ISSN 1351-3249.

Plamondon, L. (2004). L'ingénierie De La Langue Avec GATE, RALI/DIRO. Université De Montréal.

Roche, M., Heitz, T., Matte-Tailliez, O., & Kodratoff, Y. (2004). EXIT :Un Système Itératif Pour L'extraction De La Terminologie Du Domaine a Partir De Corpus Spécialisés.

Roche, M., Kodratoff, Y. (2006) Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. OTM 2006 Workshops. pp.1107-1116.

Schaffar, A., (2008). La loi Zipf est-elle pertinente? Un examen méthodologique, , XLVème colloque de l'ASRDLF, Rimouski, Canada.

Seo, C., Choi, K-S. (2007) Clustering of Wikipedia Terms with Paradigmatic Relations for Building Domain Ontology, 6th Internaltional Semantic Web Conference, 2007.

Seretan, V., (2005). Induction of syntactic collocation patterns from generic syntactic relations. In Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005), pages 1698–1699, Edinburgh, Scotland.

Shah, M. (2008). The Arabic Language. In A. Rippin, The Islamic World. (pp. 261-277). New York; London: Routledge.

Thakker, D., Sman, T., & Lakin, P. (2009). GATE JAPE Grammar Tutorial, Version 1.0. UK: PA Photos.

Zaidi, S., Laskri, M.-T., & Abdelali, A. (2010). Étude D'adaptabilité D'outils De Terminologie Textuelle à l'Arabe. COSI'10 7eme Colloque Sur L'optimisation et Les Systèmes D'information. Ouargla, Algeria.

Zipf GK. (1949). Human behaviour and principle of least effort, Cambridge, MA : Addison-Wesley.

Reusability of Quranic document using XML

Taha Zerrouki, Ammar Balla

National Computer science School,
Algiers, Algeria
taha.zerrouki@gmail.com, a_balla@esi.dz

Abstract

In this paper, we present a Quranic document modelling with XML technologies that offer the benefit of reusability. The use of this technique presents a lot of advantages in Quranic application development such as speed of development, low development cost and being readily available. Existing models fell short of combining these features. The paper review the existing models and detail the potientials of each of the models from prespectives of applications, reusability, and availibility.

1. Introduction

A large variety of Quranic applications available as desktop applications, web applications, and in other format become available for users as a concequence of advances in computer technology and its affordability. such applications provide an easier access and use of Quran. After an in-depth examinations of the area, we concluded that a dire need for a standardised specifications is more than ever. We have worked in the Quran Project, performing numerous academic studies (Zerrouki, 2006) with the goal of producing high-level models which will be used to develop standardised specifications for this large area. These standards can be used by developers and researchers to develop future applications.(Benaissa, 2008), (HadjHenni, 2009), (Toubal, 2009), for a summary, the problems in Quranic application programming can be summed in three axes:

- Othmani script: this is the most important issue of Quranic applications due to the desire to obtain an electronic typography that support identical to paper copies, by maintaining the exact marks for recitation rules. (Aldani, 1997), (Al-Suyuti, 2005),(Haralambous, 1993),(Zerrouki and Balla, 2007),(Zerrouki, 2006).

  We studied (Zerrouki and Balla, 2007) the quranic marks used across the Islamic world, used with different narrations (rewayates), we noted that the Unicode standard covers only the Hafs narration marks, and omits others marks used in North Africa and Pakistan. We proposed to add the missing marks to the Unicode standard, and we designed a Quranic font for this purpose (cf. figure 1).

- Reliability: Securing the Quran books and checking its validity is a major challenge, as the sacred text must be 100% authentic and accurate, with no tolerance for errors. This is enforced by the authorized organisms which withdraw immediately paper copies from the market when there are suspicions of errors or typos. But this is not currently available in the field of IT, given the speed of the widespread, and the inability to control things, and the absence of competent organisms which combine religion and technology (Lazrek, 2009),(Zerrouki, 2006).

  Among our extensive experiments (Zerrouki, 2006), we tried to use the numerical miracle of Quran, and studied the use of it as a mean to authenticate the sacred text. We released that is currently impossible to co-op with variance between narrations, verses counting methods and words spelling variances. If an assumption is valid for a narration, it isn't valid for another narration especially if it concerns letters and words in the absence of a standardized model.

  For the security and authentication, we thinks that is not a technique issue but is must be maintained by an authority organization which give a standard electronic version with digital signature, and white and black list of Quranic software.

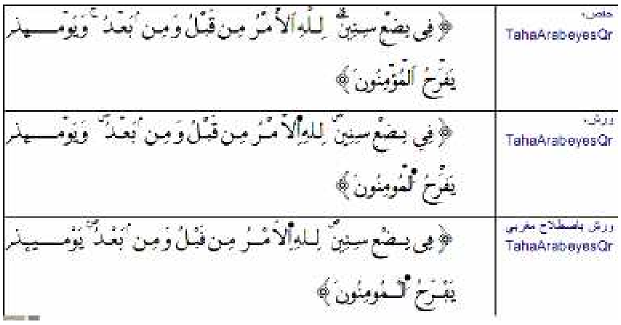


Figure 1: Quranic font supports missing quranic marks

- Exploitation of Quranic data: originally, the Quran is distributed and used in paper forms, and electronic paper-like versions (i.e. pictures). But in computing it's necessary to have it modeled, in order to manipulate this data, such as searching for a words or its meanings and to facilitate listening to a recitation, and the possibilities of memorization, etc... (Al-Suyuti, 2005),(Benaissa, 2008), (A. Bouzinne, 2006),(HadjHenni, 2009),(Toubal, 2009),(Zerrouki, 2006).

Additionally, format conversion may be fraught with risks of errors in one hand, while copyright may also stand as an obstacle to the development of such applications and utilities(Zerrouki, 2006). Therefore the urgent need to form a reference standard of Quranic applications (Zerrouki, 2006) which allow researchers to carry more productive research and produce and high quality applications.

In this paper, we'll discuss Quranic document modeling for reuse in development of various Quranic applications using XML technologies. The rest of the paper is organized as follows: We provide an extensive review about reusability and classify existing work in Quranic applications based on this criterion, we then focus on the reusable applications by presenting existing models and their pros and cons, and finally we present our own model.

2. The reusability

In computer science and software engineering, reusability is the likelihood for a segment of source code that can be used again to add new functionalities with slight or no modification. Reusable modules and classes can reduce implementation time, increase the likelihood that prior testing and use has eliminated bugs and localizes code modifications when a change in implementation is required.

Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. The technology is used to describe extended data. Among its features, it allows reusability as that is considered among the most important feature. (Attar and Chatel, 2000).

3. Classification of Quranic applications

Thousands of Quranic applications have been developed, and there are more in progress with various objectives, but we note that they are mostly closed applications and thus one cannot examine their components or study their designs. It is difficult to examine any Quranic application if it's not well documented or open source.

We can categorize Quranic applications into two important classes according to the reusability criteria: reusable applications and non-reusable, often because they are either commercial non open-source or not documented. We cannot exhaustively enumerate the applications in the non reusable category here, because most existing applications belong to this category.

4. Reusable Quranic applications

There are a limited number of reusable applications; most of which use XML technology, thus providing the ability to reuse their data in new programs or in order to extend their original functionalities. The following is a sample of these applications:

4.1. Religion 2.0 Model

This model was developed by Jon Bosak (Bosak, 1998) in 1998; and is commonly used for the Old Testament, this model is based on XML and it is composed by the following principal files:

- Tstmt.dtd: The DTD common for the four religious books (cf. figure 3)
- Bom.xml : :The book of the Mormon
- OT.xml : Old Testament (KJV)
- NT.xml : New testament (KJV)

```
<!ENTITY % plaintext "#PCDATA|i">
<!ELEMENT tstmt
   (coverpg?,titlepg?,preface?,(bookcoll|suracoll)+)>
<!ELEMENT coverpg   ((title|title2)+, (subtitle|p)*)>
<!ELEMENT titlepg   ((title|title2)+, (subtitle|p)*)>
<!ELEMENT title     (%plaintext;)*>
<!ELEMENT title2    (%plaintext;)*>
<!ELEMENT subtitle  (p)+>
<!ELEMENT preface   ((ptitle|ptitle0)+, p+,
         witlist?)+>
<!ELEMENT witlist   (witness)+>
<!ELEMENT ptitle    (%plaintext;)*>
<!ELEMENT ptitle0   (%plaintext;)*>
<!ELEMENT witness   (%plaintext;)*>
<!ELEMENT bookcoll  (book|sura)+>
<!ELEMENT book
 (bktlong, bktshort, epigraph?,bksum?,  chapter+)>
<!ELEMENT suracoll  (sura+)>
<!ELEMENT sura      (bktlong, bktshort,
              epigraph?, bksum?, v+)>
<!ELEMENT bktlong   (%plaintext;)*>
<!ELEMENT bktshort  (%plaintext;)*>
<!ELEMENT bksum     (p)+>
<!ELEMENT chapter   (chtitle, chstitle?,
          epigraph?, chsum?, (div+|v+))>
<!ELEMENT chtitle   (%plaintext;)*>
<!ELEMENT chstitle  (%plaintext;)*>
<!ELEMENT div       (divtitle, v+)>
<!ELEMENT divtitle  (%plaintext;)*>
<!ELEMENT chsum     (p)+>
<!ELEMENT epigraph  (%plaintext;)*>
<!ELEMENT p         (%plaintext;)*>
<!ELEMENT v         (%plaintext;)*>
<!ELEMENT i         (%plaintext;)*>
```

Figure 2: Bosak model DTD, tstmt.dtd

- Quran.xml : The Quran (cf. figure 3)

Evaluation of the model
Advantages

- It contains the translation of the totality of the Quran in XML

Disadvantages

- This model cannot be distributed for the Quran only (without the others books).
- The safety and authentication of Quran contents is not implemented.
- This model presents the translation of the Quran which is an application of the Quran, it don't represent the Quranic document which are expressed in Arabic language.
- The various types of Quranic fragments are not considered.
- The model is copyrighted (Bosak property), and not available without restrictions.

4.2. Arabeyes.org Model

This model is developed by Arabeyes.org (a team of open source developers), this model is designed as a goal of the Quarn Project, which gave way to Open source applications like libquran and QtQuran. The schema of this model is as follows (cf. Figure 4) (Aldani, 1997):

The following tags are used (cf. Figure 4:

```
<?xml version="1.0"?>
<!DOCTYPE tstmt SYSTEM "../common/tstmt.dtd">
<tstmt>
 <coverpg>
 <title>The Quran</title>
 <title2>One of a group of four religious
   works marked up for electronic publication
   from publicly available sources</title2>
 <subtitle>
  <p>SGML version by Jon Bosak, 1992-1994</p>
  <p>XML version by Jon Bosak, 1996-1998</p>
  <p>The XML markup and added material
     in this version are Copyright
     &#169; 1998 Jon Bosak</p>
  </subtitle>
</coverpg>
 <titlepg>
  <title>The Quran</title>
  <subtitle>
   <p>Translated by M. H. Shakir</p>
  </subtitle>
 </titlepg>
<suracoll>
  <sura>
   <bktlong>1. The Opening</bktlong>
   <bktshort>1. The Opening</bktshort>
   <v>In the name of Allah, the Beneficent,
    the Merciful.</v>
   <v>All praise is due to Allah, the Lord
    of the Worlds.</v>
   <v>The Beneficent, the Merciful.</v>
   <v>Master of the Day of Judgment.</v>
  <v>Thee do we serve and Thee do we
    beseech for help.</v>
  <v>Keep us on the right path.</v>
  <v>The path of those upon whom Thou
    hast bestowed favors. Not (the path)
    of those upon whom Thy wrath is brought
    down, nor of those who go astray.</v>
  </sura>
 </suracoll>
</tstmt>
```

Figure 3: Bosak model DTD, tstmt.dtd

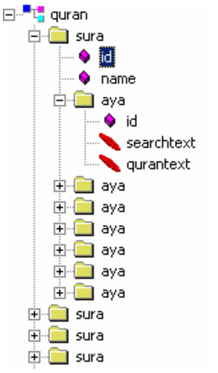
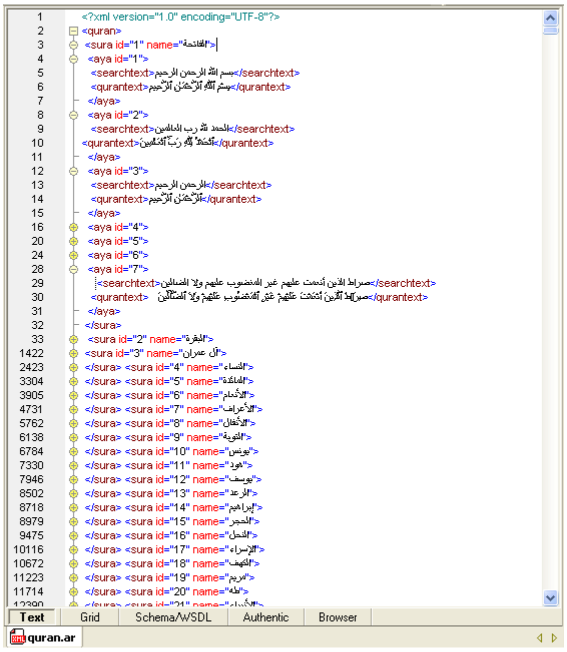


Figure 4: Diagram XML of Arabeyes.org Quran Project.

- <quran>: which includes the complete content of the Quran.
- <sura>: a sura of the Quran. The attribute name is the name of sura. The attribute id is the ordinal number of the sura.
- <aya>: an aya, which has an id (the aya number in the sura), <searchtext>and <qurantext>elements.
- <searchtext>: Quranic text without vowels, used for research.
- <qurantext>: Quranic text with vowels.



```
<?xml version="1.0" encoding="UTF-8"?>
<quran>
 <sura id="1" name="الفاتحة">
  <aya id="1">
   <searchtext>بسم الله الرحمن الرحيم</searchtext>
   <qurantext>بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ</qurantext>
  </aya>
  <aya id="2">
   <searchtext>الحمد لله رب العالمين</searchtext>
<qurantext>ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ</qurantext>
  </aya>
  <aya id="3">
   <searchtext>الرحمن الرحيم</searchtext>
   <qurantext>ٱلرَّحْمَٰنِ ٱلرَّحِيمِ</qurantext>
  </aya>
  <aya id="4">
  <aya id="5">
  <aya id="6">
  <aya id="7">
    <searchtext>صراط الذين أنعمت عليهم غير المغضوب عليهم ولا الضالين</searchtext>
   <qurantext>   صِرَٰطَ ٱلَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ ٱلْمَغْضُوبِ عَلَيْهِمْ وَلَا ٱلضَّآلِّينَ</qurantext>
  </aya>
 </sura>
  <sura id="2" name="البقرة">
 <sura id="3" name="آل عمران">
 </sura> <sura id="4" name="النساء">
 </sura> <sura id="5" name="المائدة">
 </sura> <sura id="6" name="الأنعام">
 </sura> <sura id="7" name="الأعراف">
 </sura> <sura id="8" name="الأنفال">
 </sura> <sura id="9" name="التوبة">
 </sura> <sura id="10" name="يونس">
 </sura> <sura id="11" name="هود">
 </sura> <sura id="12" name="يوسف">
 </sura> <sura id="13" name="الرعد">
 </sura> <sura id="14" name="إبراهيم">
 </sura> <sura id="15" name="الحجر">
 </sura> <sura id="16" name="النحل">
 </sura> <sura id="17" name="الإسراء">
 </sura> <sura id="18" name="الكهف">
 </sura> <sura id="19" name="مريم">
 </sura> <sura id="20" name="طه">
 </sura> <sura id="21" name="الأنبياء">
```

Figure 5: Fragment from arabeyes xml file.

Evaluation of the model

Advantages

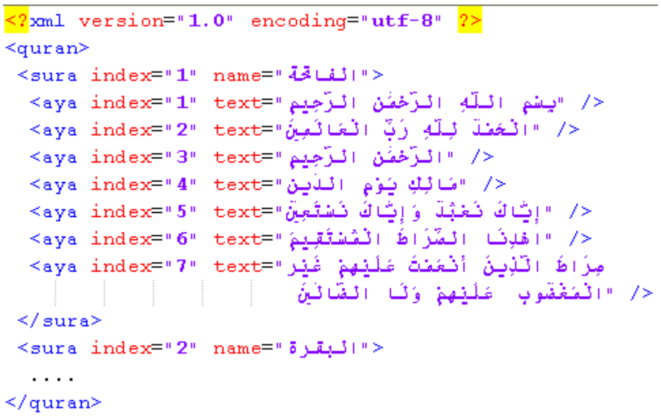
- Open source: model not proprietary, code available for free access;
- This XML model contains the text of the totality of the Quran in Arabic language (Fig. 5).
- The Othmani script is supported.
- Model based on Unicode coding.

Disadvantages

- Modeling of only one type of Quran fragments: in ayates and sourates;
- Security and authentication of the Quranic contents are not implemented;
- No information about the Quranic book (rewayates, writing rules, fragments, version, committee of validation, etc.);
- Oversize of the file quran.xml, due to the double writing of each aya to present Quranic text without vowels, whereas it can be obtained programmatically;
- Some Quranic marks are not represented in Unicode.

4.3. Tanzeel Model

Tanzil is a Quranic project launched early 2007 to produce a highly verified precise Quran text to be used in Quranic websites and applications. The mission of the Tanzil project is to produce a standard Quran text and serve as a reliable source for this standard text on the web. The Tanzil project has integrated several previous works done in a number of Quranic projects to obtain a unified quran text, and brought



```
<?xml version="1.0" encoding="utf-8" ?>
<quran>
 <sura index="1" name="الفاتحة">
  <aya index="1" text="بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ" />
  <aya index="2" text="الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ" />
  <aya index="3" text="الرَّحْمَٰنِ الرَّحِيمِ" />
  <aya index="4" text="مَالِكِ يَوْمِ الدِّينِ" />
  <aya index="5" text="إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ" />
  <aya index="6" text="اهْدِنَا الصِّرَاطَ الْمُسْتَقِيمَ" />
  <aya index="7" text="صِرَاطَ الَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ
                        الْمَغْضُوبِ عَلَيْهِمْ وَلَا الضَّالِّينَ" />
 </sura>
 <sura index="2" name="البقرة">
  ....
</quran>
```

Figure 6: Tanzil project: quran-simple.xml.

that text to a high level of accuracy by passing it through several manual and automatic verification phases.
The Tanzil project has two XML documents, one for the Quran text and the second for the Quran metadata, which contains metadata about the Quran structure and fragments. The Quran xml document has multiple versions according to the typography used: simple with the Arabic standard script (Fig. 6), or Uthmani script (which need a specific font to display text). The structure of the Quran xml document is basic: it contains Sura (Chapter) elements which in turn contain aya (verse) elements. Each sura has a name and an index, and each aya has an index and a text content(Zarrabi-Zadeh, 2010).
The metadata xml document contains information about the Mushaf (Holy Book) structure:

- A suras (chapters) element which groups sura elements. Each sura element has information about Number of ayas 'ayas', the start aya number 'start', an Arabic name 'name', translated name 'tname', English name 'ename', revelation place 'type', order of revelation 'order' and order number of rukus 'rukus'.
- A juzs (part) element which groups juz elements. Each juz element has attributes: index number 'index', sura index of start 'sura', and aya index 'aya'.
- A hizbs (group) element which groups quarters elements. Each quarter element has attributes: index number 'index', sura index of start 'sura', and aya index 'aya'
- A manzils (station) element which groups manzil elements. Each manzil element has attributes: index number 'index', sura index of start 'sura', and aya index 'aya'
- A rukus (section) element which groups ruku elements. Each ruku element has attributes: index number 'index', sura index of start 'sura', and aya index 'aya'
- A pages (page number in Medina Mushaf) element which groups page elements. Each page element has attributes: index number 'index', sura index of start 'sura', and aya index 'aya'

```
<?xml version="1.0" encoding="utf-8" ?>
<quran type="metadata" version="1.0"
 copyright="(C) 2008-2009 Tanzil.info" license="cc-by">
<suras alias="chapters">
<sura index="1" ayas="7" start="0"
 name="" tname="Al-Faatiha"
 ename="The Opening" type="Meccan"
 order="5" rukus="1" />
<sura index="2" ayas="286" start="7"
  name="" tname="Al-Baqara"
  ename="The Cow" type="Medinan"
  order="87" rukus="40" />
...
</suras>
<juzs alias="parts">
<juz index="1" sura="1" aya="1" />
<juz index="2" sura="2" aya="142" />
...
</juzs>
<hizbs alias="groups">
<quarter index="1" sura="1" aya="1" />
<quarter index="2" sura="2" aya="26" />
...
</hizbs>
<manzils alias="stations">
<manzil index="1" sura="1" aya="1" />
<manzil index="2" sura="5" aya="1" />
...
</manzils>
<rukus alias="sections">
<ruku index="1" sura="1" aya="1" />
<ruku index="2" sura="2" aya="1" />
...
</rukus>
<pages>
<page index="1" sura="1" aya="1" />
<page index="2" sura="2" aya="1" />
...
</pages>
<sajdas>
<sajda index="1" sura="7"
  aya="206" type="recommended" />
...
<sajda index="15" sura="96"
   aya="19" type="obligatory" />
</sajdas>
</quran>
```

Figure 7: Tanzil Project: quran-data.xml.

- A sajdas (location of postration) element which groups sajda elements. Each sajda element has attributes: index number 'index', sura index of start 'sura', and aya index 'aya', and a type of sajda ('recommended or obligatory).

Evaluation of the model

Advantages

- Open source.
- The Othmani script is supported.
- Model based on Unicode coding.
- Uses XML to describe Quran properties.
- Vocalized text.
- Modeling of various types of Quran fragments: hizbs, manzils, rukus, pages;

Disadvantages

- Security and authentication of the Quranic contents are not implemented;

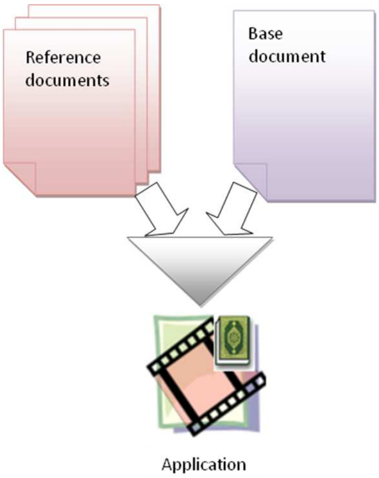


Figure 8: The diagram of electronic Mushaf.

- No information about the Quranic books (publisher, rewayates, writing rules, version, committee of validation, etc.);.

5. Proposed model

Our idea is to separate the base document from others referential information which are added to the Mushaf by scholars in order to facilitate the Quran learning and recitation. This method is adopted by Tanzeel project also, but we add more information about the electronic document, and about quranic text like books used as reference for fragmentations and ayates count. The added information support more narrations and typography conventions which differ from Islamic region to another.

As mentioned above, the authentication of Quranic document is a core issue, then we provide more elements and information about the organisms which give licenses or authorization to publish a new document or application. We think that the only way to protect Quran and ensure its integrity is to assume responsibility by an authorized organization which can give a numeric signature.

In other hand, our model attempts to provide more information about fragmentation conventions which are not used in Tanzeel, by handling different conventions used in different regions of Islamic world. For example, the ruku' and manzil fragmentation are used only in Indian world. The thumun fragmentation is used in the North Africa. The Tanzeel project handle Juz', Hizb, Rub', Ruku', Manzil, but ignore the Thumun fragmentation used in North Africa.

In this model, we propose the general outline of an electronic document Model of the Mushaf. This diagram is built around (Fig. 8) (Zerrouki, 2006):

- **A Basic document**: contains the Quran text organized in Chapters and verses.
- **A set of reference documents**: these documents contain additional information about the Mushaf, such as information about publication, fragments, the positions of waqfs (punctuation like , a reference to indicate sentences pauses in the Quranic text), etc.

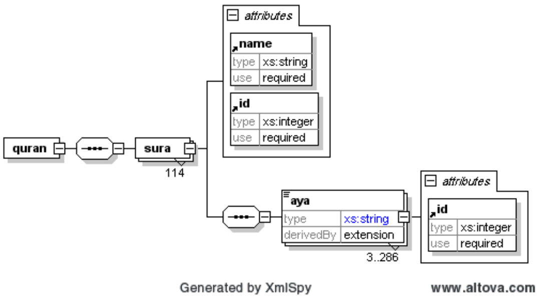


Figure 9: XML schema of quran basic document.

```
<xs:element name="Quran">
 <xs:complexType >
  <xs:sequence >
   <xs:element name="will sura"
     minOccurs="114" maxOccurs="114">
    <xs:complexType>
     <xs:sequence>
      <xs:element name="aya"
        minOccurs="3" maxOccurs="286">
       <xs:complexType >
        <xs:simpleContent >
         <xs:extension base="xs:string">
          <xs:attribute ref.="id"
                  use="required"/>
         </xs:extension>
        </xs:simpleContent>
       </xs:complexType>
      </xs:element>
     </xs:sequence>
     <xs:attribute ref.="name"
               use="required"/>
     <xs:attribute ref.="id"
                use="required"/>
    </xs:complexType>
   </xs:element>
  </xs:sequence>
 </xs:complexType>
</xs:element>
```

Figure 10: XML Schema of basic document.

5.1. The base document

The basic document contains only the Quran text, which is organized in chapters and verses. The Quran contains exactly 114 chapters, each chapter has a unique name, an ordinal number (id), and a set of ayates. The number of ayates per sura is between 3 and 286 (AL-Aoufi, 2001), (Mus'haf, 2000), (Mus'haf, 1966), (Mus'haf, 1998), (Mus'haf, 1987) . The XML schema of the basic document is illustrated in Fig. 9.

5.2. Reference documents

The reference documents contain different references about the basic documents, these data are not a Quranic text, but they are necessary for the use of Quran. These references are Mushaf Properties, Fragments, Surates revelation places and Sajdahs.

The XML schema joins the different sub-schema in one reference.

Next, we describe the different reference document in details.

5.2.1. Mushaf properties document

This document contains the properties of an quran book (Mushaf). We can classify properties in two categories (cf. Fig. 12) :

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
   elementFormDefault="qualified"
   attributeFormDefault="unqualified">
 <xs:include schemaLocation="ref_Mushaf.xsd"/>
 <xs:include  schemaLocation="fragementation.xsd"/>
 <xs:include schemaLocation="reve_nb_aya.xsd"/>
 <xs:include schemaLocation="sajdahs.xsd"/>
</xs:schema>
```

Figure 11: XML Schema of Quran reference documents.

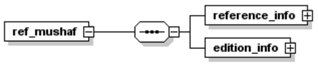


Figure 12: XML Schema of Mushaf properties document.

- Information about the quranic text.
- Information about Mushaf edition.

5.2.2. Information about the Quran text

This information allows describing the references used to write this Mushaf, when we went to edit and publish a new Mushaf we must specify some properties and references (AL-Aoufi, 2001), (Mus'haf, 2000), (Mus'haf, 1966), (Mus'haf, 1998), (Mus'haf, 1987).

1. Rewayat: describe the chain of narrators of the Quran from the prophet Mohamed.
2. Spelling (Hijaa): describe the base document, to rewrite the Mushaf, it is the Othman Mushaf named Al-Imam, all Mushafs must be written according to the Imam Mushaf,
3. Diacritization : indicate the type of the diacritics used on Quran text, like conventions used in North Africa, or middle-east.
4. Ayates counting method: there is some difference in the ayate counting, this information must be used to indicate which method is used.
5. Typography conventions: describe the rules used in the typography of the Quran text, this means all symbols used for waqfs, specials delimiters and recitations marks..
6. Fragmentation: This element is used to indicate the fragmentation method and its reference. The detailed fragmentation information is illustrated in a separated document (fragmentation document).
7. The Revelation place: there are some chapters and verses reveled in Mecca, and others reveled in Medina, The indication of revelation places helps in the interpretation of the Quranic text. This element is used to indicate the reference of the revelation information. The detailed revelation places information is illustrated in a separated document (revelation document).
8. Indication of waqfs : this element indicated the type of waqfs used, and the reference of waqfs.

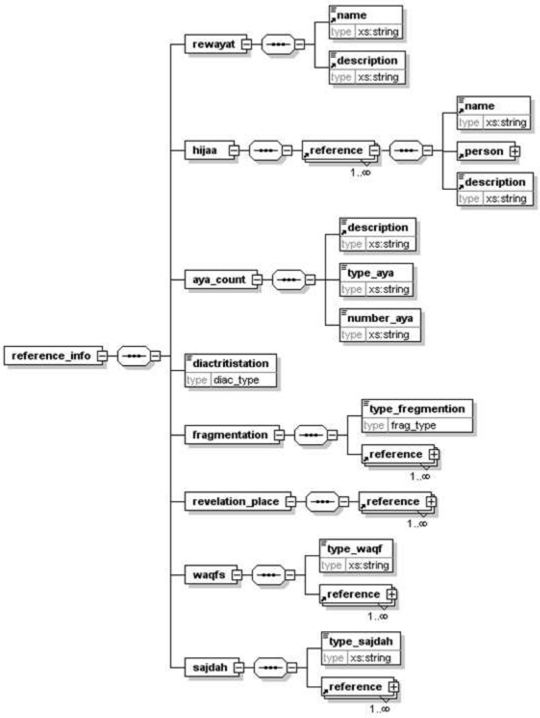


Figure 13: XML Schema of information about the Quran text.

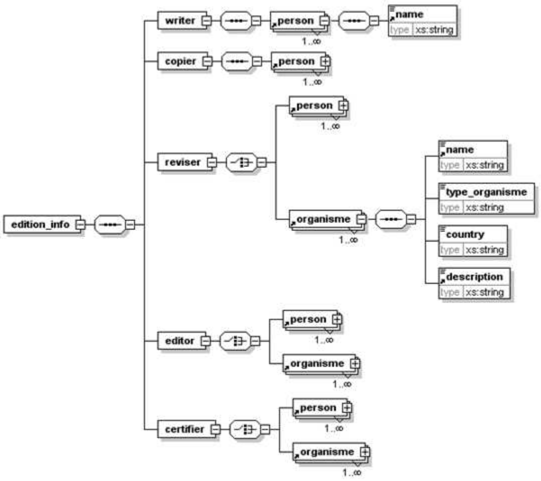


Figure 14: XML Schema of information about the Mushaf edition.

9. Indication of sajdahs : this element indicated the type of sajdahs used, and the reference of sajdahs.

5.2.3. Information about Mushaf edition

This information is about the edition of Mushaf, it contains references of the writer of the text, the copier from a paper Mushaf, the review committee, the publisher, and the certifier organisation.

5.2.4. Fragmentation document

The Quran text is originally fragmented into surates and ayates, but there are other fragmentations added in order to facilitate the recitation. The whole Mushaf is fragmented

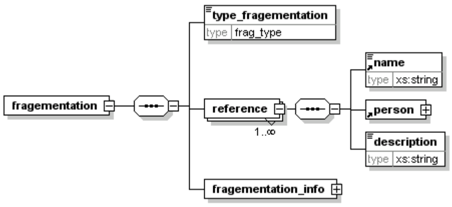


Figure 15: details of fragmentation_info.

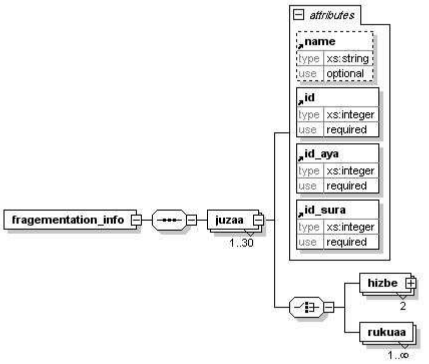


Figure 16: details of fragmentation_info juz'.

into 30 parts named juz'. Each juz' is fragmented into 2 sub-parts named hizb. Each hizb has sub-parts (Nisf:1/2 hizb, rub': 1/4 hizb, Thumn: 1/8 hizb). The Thumn is used only in North Africa. (AL-Aoufi, 2001), (Mus'haf, 2000), (Mus'haf, 1966), (Mus'haf, 1998), (Mus'haf, 1987).

There are other fragmentations like Manzel which represent the 1/7 of the whole Mushaf. The Ruku' is a part of a sura which contains one topic and can be read in prayer (AL-Aoufi, 2001).

The fragmentation document contains detailed information about fragmentation of the Mushaf. The root element contains:

- Fragmentation type: thuman (north Africa), rub' (middle-east),ruku'(indian world).
- The reference of fragmentation: the scholar book.
- The detailed fragmentation information (cf. Fig. **??**, Fig. 16) about juz' and hizb fragmentation, each fragment has an ordinal number as identifier, and a start aya an a start sura number (cf. Fig. 17).

5.2.5. Revelation document

This document describes the revelation places of surates, the root element contains surates elements, each sura element has a revelation place and a number of its ayates (cf. 18)

5.2.6. Sajdahs Document

This document describe the places of the sajdah in the Mushaf. Each place is located by the aya number and the sura number (cf. Fig. 19).

Evaluation of the model

Advantages

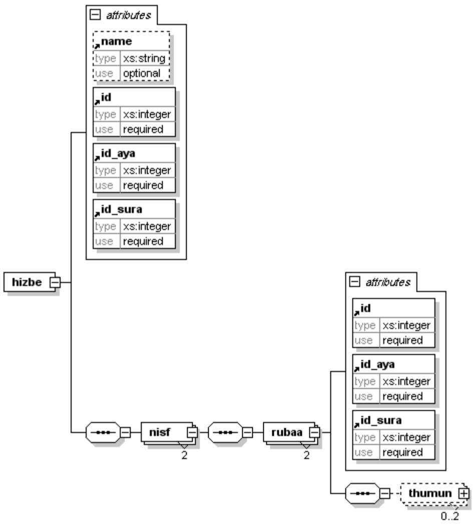


Figure 17: Hizb Fragmentation.

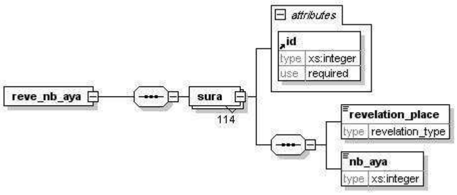


Figure 18: Revelation document structure.

- Open source: model not proprietary, code available for free access;
- This XML model contains the totality of the text of the Quran in Arabic language.
- The Othmani script is supported.
- This model implements all the possible types of fragmentation (Ahzab, Rukaa, etc.)
- Complete description of Mushaf (rewayates, writing rules, fragmentations, version, committee of validation, etc.);
- Security and authentication of the contents of the Quran exists but not yet implemented.

Disadvantages

- This model is multi structured; we cannot merge these entire Quran XML schema in one only XML schema.

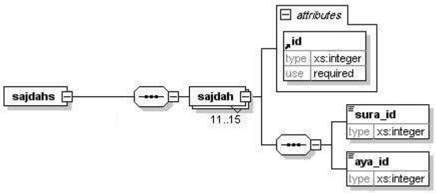


Figure 19: Sajdahs document structure.

Criteria	Religion 2.0	arabeyes	Tanzil	Our model
Open source		+	+	+
Vocalized text		+	+	+
Othmani script		+	+	+
Unicode		+	+	+
Mushaf information			+	++
Reusability		+	+	+
Mushaf fragmentation			+	++
XML Document	+	+	+	++
Quranic sciences	+			+
Translations		+	+	+

Table 1: Comparison between models

5.2.7. Comparison between models

We can summarize comparison between previous models in this table:
In other hand, every copy or published mush'afs must mention information about the publisher and the writer and the certifying entity, in order to prevent errors and assume responsibilities in case of errors.
We detailed this special features in :

1. **information about the Quranic text**: The reference document provides more information about the given Quranic text by specifying the scholar books used to verify every letter and every word spelling. It gives also information about the rewayate, the aya counting method. All this information affect the structure of the text and its authenticity.

   Regarding the waqf markers, in Maghreb, they use a simple mark of waqf as a small high Sad letter, this kind of waqf is called Waqf Mihbati, according to the reference book used to specify the position of each waqf in the text. But in middle east, they use different kind of waqfs with different degree of recommendation (obligatory, recommended, forbidden,).

   Our model provides information about waqfs as separate elements on the text, while Tanzeel represent it as marks in the text.

2. **Information about mus'haf edition**: Provide more references about the writer, editor and copier of the mus'haf. The writer is the person who hand-writes the Quran text in printed mushaf like Uthman Taha; The editor is a person or an organism which publish this Mus'haf; the copier is the person who copies the electronic document from printed mus'haf. It can be used to authenticate the mus'haf and assume responsibilities of every party in order to assure accuracy of the text. Our model supports different conventions of fragmentation based on juz', hizb, rub', thumn, ruku' and manzil used across the Islamic regions. Tanzeel project doen't support thumn fragmentation.!!!!

We have tried in our model to avoid deficiencies found in other models, particularly in the area reusability which is the main aim of this project, and not the Quranic application development.

6. Other works

This model helps us provide fundamental API for developers to develop best applications and libraries about Quran in an open source context. Recently, we have developed Alfanous, a Quranic search engine that offers simple and advanced search services for information contains in Holy Quran What information??? Specify some-. It is based on the modern approach of information retrieval to get a good-stability??? and a fast retrieval (Chelli et al., 2011).
The idea behind the project is to collect all features developed in other projects in one open source project, to allow developers to produce high-quality applications for Quran.

7. Conclusion

Our project aims to create a standard specification model using advanced technologies such as XML and ontologies in order to serve Quran by the best methods available. Through this study, we reviewed existing technology used to produce Quran software with focus on reusability. The various formats and standards bring the need to develop a comprehensive model that captures all the requirements and supports all the features. Such solution would reduce the development time and cost and reduce potential errors, for which such applications are very sensitive. For future work, we are expanding our model to build Ontological representation of Quran. The new model will allow semantic representation which will be used to facilitate information retrieval from Quran.

8. References

M. Remadhna A. Bouzinne. 2006. Xml model for quranic applications.

M.S AL-Aoufi. 2001. The progress of writing mushaf. Technical report, he King Fahad Complex for Printing of the Holy Quran, Saudi Arabia.

J. Al-Suyuti. 2005. *Quranic sciences', 'al-Itqan fi'Ulum al-Qur'an'*. Dar alghad aljadid, Cairo, Egypt.

Abu Amr Aldani. 1997. *The Mushafs diacritization 'Al-muhkem fi naqt al-masahif'*. Dar- Al-fikr, Syria.

P. Attar and B. Chatel. 2000. *State of recommendations XML, in the documentary field.* Books GUTenberg.

K. Benaissa. 2008. Reusable and standard model for applications around quran. Master's thesis, National School of Computing , Algiers, Algeria.

J. Bosak. 1998. The religion 2.0. Technical report, .

A. Chelli, T. Zerrouki, A. Balla, and Dahmani M. 2011. Improving search in holy quran by using indexes. In *Third National Information Technology Symposium Arabic and Islamic Contents on the Internet*, Riyadh , Saudi Arabia. The College of Computer and Information Sciences.

M. HadjHenni. 2009. Semantic modelling, indexation et query of quranic document: using ontologies. Master's thesis, National School of Computing , Algiers, Algeria.

Y. Haralambous. 1993. Typesetting the holy quran with tex. In *TeX and arabic script*, Paris.

A. Lazrek. 2009. Standard electronic structure holy quran and digital authentication and certification automation. In *International Conference about "The Glorious Qur'an and Information Technology*, Al-madina, Saudi Arabia. King Fahad Complex for Printing of the Holy Quran.

Mus'haf. 1966. *Mushaf in rewayat of Warch*. Abd Errahman Mohamed edition, Egypt.

Mus'haf. 1987. *Mushaf in rewayat of Warch*. Dar Al-fikr , Syria.

Mus'haf. 1998. *Mushaf in rewayat of Warch*. Ministry of Religious Affairs, Algeria.

Mus'haf. 2000. *Mushaf al-Madinah an-Nabawiyyah, in the rewayat of Hafs*. King Fahad Complex for Printing of the Holy Quran, Madina, Saudi Arabia.

E. Toubal. 2009. Quran recitation rules modelling : using ontologies. Master's thesis, National School of Computing , Algiers, Algeria.

H. Zarrabi-Zadeh. 2010. Tanzil project. Technical report, Tanzio Project http://tanzil.info.

T. Zerrouki and A. Balla. 2007. Study of electronic publishing of holy quran. In *First International Symposium On Computer and Arabic Language*, Ryadh, Saudi Arabia. King Abdulaziz City for Science and Technology.

T. Zerrouki. 2006. Toward a standard for electronic document of quran. Master's thesis, National School of Computing.