LRE-REL2

Proceedings of the 2nd Workshop on

Language Resources and Evaluation for Religious Texts

31 May 2014, Reykjavik, Iceland

A post-conference workshop co-hosted at LREC'2014
Language Resources and Evaluation Conference

Editors:

Claire Brierley, Majdi Sawalha, Eric Atwell

2nd Workshop on Language Resources and Evaluation for Religious Texts

Workshop Programme

31 May 2014

14:00 – 16:00 Session 1 Papers
14:00 Claire Brierley and Majdi Sawalha (Workshop Chairs)
Introduction to the 2nd Workshop on Language Resources and Evaluation for Religious Texts
14.10 Plenary Speaker: Majdi Sawalha (with co-authors Claire Brierley and Eric Atwell)
Automatically-generated, phonemic Arabic-IPA Pronunciation Tiers for the Boundary-Annotated Qur'an Dataset for Machine Learning (version 2.0)
14.40 Claudia Resch, Thierry Declerck, Barbara Krautgartner, and Ulrike Czeitschner
ABaC:us Revisited - Extracting and Linking Lexical Data from a Historical Corpus of Sacred Literature
14.50 Bruno Bisceglia, Rita Calabrese, Ljubica Leone
Combining Critical Discourse Analysis and NLP Tools in Investigations of Religious Prose
15.00 Daniela Gîfu, Liviu-Andrei Scutelnicu and Dan Cristea
Humour and Non-Humour in Religious Discourse
15.10 Manal AlMaayah, Majdi Sawalha and Mohammad A.M. Abushariah
A Proposed Model for Qur'anic Arabic Wordnet
15.20 Sameer M. Alrehaili and Eric Atwell
Computational Ontologies for Semantic Tagging of the Qur'an: A Survey of Past Approaches
15.30 Ahmad Alqurneh and Aida Mustapha
Traditional vs. Chronologic*al Order: Stylistic Distance Analysis in Juz' Amma*
15.40 Kamal Abou Mikhael
The Greek-Arabic New Testament Interlinear Process: greekarabicnt.org

16:00 – 16:30 Coffee break

16:30 – 17:15 Session 2: Posters, Discussion, and Networking
Poster presentations from all authors listed above

17:15 – 18:00 Session 3: Plenary Discussion led by Workshop Chairs
Topic: Research agenda for LRE-Rel

18:00 End of Workshop

Editors

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

Workshop Organizers/Organizing Committee

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

Workshop Programme Committee

Muhammad A.M. Abushariah	Computer Information Systems, University of Jordan, Jordan
Eric Atwell	School of Computing, University of Leeds, UK
Claire Brierley	School of Computing, University of Leeds, UK
Liviu Dinu	Centre for Computational Linguistics, University of Bucharest, Romania
Kais Dukes	School of Computing, University of Leeds, UK
Moshe Koppel	Department of Linguistics, University of Jordan, Jordan
Dag Haug	Department of Philosophy, History of Art and Ideas, University of Oslo, Norway
John Lee	Halliday Centre for Intelligent Applications of Language Studies, City University of Hong Kong, (HK)
Deryle Lonsdale	Department of Linguistics and English Language, Brigham Young University, US
Bassam Hammo	Department of Computer Science, Bar-Ilan University, Israel
Bob MacDonald	Research and Development, Anthony Macauley Associates, Canada
Sane Yagi	Mathematics and Computer Science, Mohammed 1st University, Morocco
Mohamed Menacer	Taibah University, Saudi Arabia
Behrooz Minaei	School of Computer Engineering, Iran University of Science and Technology, Iran
Aida Mustapha	Department of Computer Science and Information Technology, Putra University, Malaysia
Nadeem Obaid	Computer Information Systems, University of Jordan, Jordan
Nils Reiter	Department of Computational Linguistics, Heidelberg University, Germany
Claudia Resch	Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences, Austria
Mortaza Rezaee	Islamic College, London, UK
Majdi Sawalha	Computer Information Systems, University of Jordan, Jordan
Gurpreet Singh	Centre for Language and Communication Studies, Trinity College Dublin, Ireland
Janet Watson	Arabic and Middle Eastern Studies, and Linguistics and Phonetics, University of Leeds, UK
Andrew Wilson	Department of Linguistics and English Language, University of Lancaster, UK
Azzeddine Mazroui	Computer Information Systems, University of Jordan, Jordan

Table of contents

Author Index

Language Resources and Evaluation for Religious Texts 2: Introduction

Claire Brierley, Majdi Sawalha and Eric Atwell

Welcome to LRE-Rel2, the second Workshop on Language Resources and Evaluation for Religious Texts. After a successful launch at LREC 2012 in Istanbul, Turkey, we have organised this second workshop hosted by LREC 2014 in Reykjavik, Iceland. This is an inclusive workshop, aimed at researchers with a generic interest in religious texts to raise awareness of different perspectives and practices, and to identify some common themes. Our first workshop attracted a range of scholarship, particularly on Arabic and Islamic Studies, and this year we were keen to extend this range to canonical texts from other languages and religions, and to foster inter-faith corpus studies, tracing similarities as well as differences in religious texts, where this genre includes: the faith-defining religious canon; authoritative interpretations and commentary; sermons, liturgy, prayers, poetry, and lyrics.

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

- measuring semantic relatedness between multiple religious texts and corpora from different religions;
- 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;
- 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-mediated methods for analysing religious text

Our own research has focussed on Arabic Natural Language Processing, and in particular, Qur'anic Arabic (cf. our papers in the LRE-Rel1 and LRE-Rel2 Workshops and main LREC*'2010/12/14* Conference Proceedings); but we were pleased to receive papers dealing with a range of other holy books and religious texts, both historical and contemporary, with an interesting focus this year on the vernacular and on register e.g. historical German (1650-1750), and manifestations of humour in Romanian sermons. Many of the papers present an analysis technique applied to a specific religious text, which could also be relevant to analysis of other texts, including: automated, IPA-based transcription; specification of search patterns via regular expressions; stylometry, and detecting similarities and correspondences between texts; text extraction; semantic annotation and modelling; genre and critical discourse analysis. As an innovation this year, we will seek to identify a common research agenda for LRE-Rel during the plenary session.

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.

References

M AlMaayah, M Sawalha and M Abushariah. 2014. A Proposed Model for Qur'anic Arabic Wordnet. Proc LRE-Rel2, Reykjavik, Iceland.

S Alrehaili and E Atwell. 2014. Computational Ontologies for Semantic Tagging of the Qur'an: A Survey of Past Approaches. Proc LRE-Rel2, Reykjavik, Iceland.

E Atwell, C Brierley and M Sawalha. 2012. Introduction to Language Resources and Evaluation for Religious Texts. Proc LRE-Rel1, Istanbul, Turkey.

C Brierley, M Sawalha and EAtwell. 2014. Language Resources and Evaluation for Religious Texts 2: Introduction. Proc LRE-Rel2, Reykjavik, Iceland.

C Brierley, M Sawalha and E Atwell. 2014. Tools for Arabic Natural Language Processing: a case study in qalqalah prosody. Proc LREC’2014, Reykjavik, Iceland.

C Brierley, M Sawalha and E Atwell. 2012. Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing. Proc LREC'2012, Istanbul, Turkey.

K Dukes and E Atwell. 2012. LAMP: A Multimodal Web Platform for Collaborative Linguistic Analysis. Proc LREC'2012, Istanbul, Turkey.

K Dukes, E Atwell and A Sharaf. 2010. Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. Proc LREC'2010. Valletta, Malta.

M Sawalha, C Brierley and E Atwell. 2014. Automatically-generated, phonemic Arabic-IPA Pronunciation Tiers for the Boundary-Annotated Qur'an Dataset for Machine Learning (version 2.0). Proc LRE-Rel2, Reykjavik, Iceland.

M Sawalha, C Brierley and E Atwell. 2012. Predicting Phrase Breaks in Classical and Modern Standard Arabic Text. Proc LREC'2012, Istanbul, Turkey.

M Sawalha and E Atwell. 2010. Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text. Proc LREC'2010. Valletta, Malta.

A Sharaf and E Atwell. 2012. QurAna: Corpus of the Quran annotated with Pronominal Anaphora. Proc LREC'2012, Istanbul, Turkey.

A Sharaf and E Atwell. 2012. QurSim: A corpus for evaluation of relatedness in short texts. Proc LREC'2012, Istanbul, Turkey.

The Greek-Arabic New Testament Interlinear Process: `greekarabicnt.org`

Kamal Abou Mikhael
American University of Beirut
Beirut, Lebanon
kamal@greekarabicnt.org

Abstract

Greek-Arabic New Testament (NT) interlinears have yet to be published in a digital text format. Moreover, a corpus-based approach to the evaluation of Arabic translations of the NT has yet to be employed. The resource needed to actualize both of these goals is an aligned database of the source text and translation. This paper describes a process for creating such a database that combines automation and manual editing to create the database from which interlinears and reverse interlinears can be generated. The approach is text-based and leverages the text processing commands, editors, and scripting languages available in the **nix* environment and uses freely available linguistic tools and NT texts. The paper also discusses issues of Arabic morphology and orthography that must be addressed in the preparation and alignment of the data. It is based on an initiative by `greekarabicnt.org` to align the Smith-Van Dyck Arabic Translation of the NT with its source text, Mill's Textus Receptus. At the time of writing, Chapters 1 and 2 of the First Epistle of John have been aligned, and interlinear and reverse-interlinear prototypes have been created in electronic format.

Keywords: Greek New Testament, Arabic, Interlinear

1. Introduction

Greek-English New Testament (NT) interlinears (and reverse interlinears) are widely available and serve as a valuable reference tool for students of the Greek NT. Regardless of whether one rejects or accepts the NT as true, wholly or in part, an interlinear can help facilitate an understanding that considers the original language of the NT, which makes subsequent discussion more meaningful.

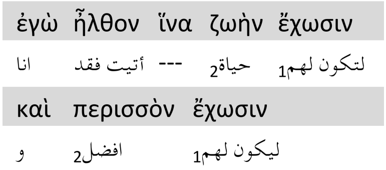
However, Greek-Arabic interlinears have yet to be published in a digital text format. Beyond a visual product, what is also desired is a database of the alignment of source and translation text that would facilitate a corpus-based study and evaluation of a given translation. Creating a resource that caters to both of these goals is feasible due to the free availability of fully lemmatized and annotated Greek NT digital texts as well as Arabic NT translations. In addition to data sources, freely available text processing and linguistic tools/resources can be harnessed to process the source texts and prepare them for the alignment process.

The organization of the paper is as follows. Section 2 presents basic terms and concepts related to interlinears as well as the aspects of the Greek and Arabic languages that are pertinent to the project. Section 3 covers related works on the Greek NT that are either related to the Arabic language or the methodology described in this paper. Section 4 outlines the process identified for creating the Greek-Arabic interlinears and describes the available resources and tools. It breaks down the process into four phases and corresponding subsections related to preparing the Greek text (Section 4.1), preparing the Arabic text (Section 4.2), aligning the two texts (Section 4.3), and generating a displayable format (Section 4.4). Section 5 discusses findings related to the Greek source text, the processing of the Arabic text, and the alignment of the two texts. The findings are based on the text aligned so far, Chapters 1 and 2 of the First Epistle of John (hereafter, I John 1–2). Section 6 discusses enhancements and useful additions to the project as well as types of analysis facilitated by the aligned database that is being created. The conclusion briefly reflects on the goals of the project and significance of its final product.

2. Background

This section defines key terms, concepts, and issues related to the project. Due to constraints in space, this section will not cover common knowledge regarding the Arabic language, the content of the NT, and basic linguistic terms. However, to ensure a proper understanding of examples, the reader must note that Arabic is read from right to left.

Interlinears contain the alignment of a source text and a translation, where the word order of the source text is preserved (Figure 1). Reverse interlinears, on the other hand, preserve the word order of the translation (Figure 2). In addition to the source text and translation, an interlinear can contain annotations such as lemmas, parsing information, and one or more glosses in the translation language. Hereafter, the term *interlinear* refers to both interlinears and reverse interlinears unless otherwise stated.



ἐγὼ	ἦλθον	ἵνα	ζωὴν	ἔχωσιν
انا	أتيت فقد	---	حياة$_2$	لتكون لهم$_1$
καὶ	περισσὸν	ἔχωσιν		
و	افضل$_2$	ليكون لهم$_1$		

Figure 1: Greek-Arabic interlinear text.

In addition to defining an interlinear and its component parts, this section must disambiguate the terms *Greek* and *Arabic*. At once they refer to a particular text and a particular language. Various Greek NT texts exist and they differ depending on the textual sources used and the methodology employed to choose between variant readings

حياة	لتكون لهم	---	فقدأتيت	انا
Ζωὴν	ἔχωσιν	ἵνα	ἦλθον	ἐγὼ
		افضل	ليكون لهم	و
		περισσὸν	ἔχωσιν	καὶ

Figure 2: Arabic-Greek reverse interlinear text.

(Porter, 2013)[1]. However, all use the same language, Koine Greek (hereafter referred to as Greek or Biblical). In addition, various Arabic translations exist and they differ depending on the Greek text(s) they used as well as the type of language used and underlying approach to translation (e.g., formal equivalence vs. dynamic equivalence). The Arabic translation used in this project is the Smith-Van Dyck translation (svd, 1865), which was completed by Protestant missionaries in 1865 in collaboration with Arab. It is considered a literal translation and it lies between Classical Arabic and Modern Standard Arabic (Smith, 1900). The NT of the SVD in based on the Textus Receptus[2] published by Mill (Hall, 1885).

One issue particular to Greek-Arabic interlinears is that many morphemes in an Arabic word can correspond to more than one word in the Greek text. The definite article, prepositions, and conjunctions are prefixes that are concatenated to the words that follow them. Similarly, accusative and genitive pronouns are suffixed to the words they modify. On the other hand, there are distinct Arabic morphemes such as plural markers and verbal inflections indicating person/number that do not need to be treated as separate morphemes because the corresponding individual Greek verbs and nouns are inflected with corresponding information.

Consider the word واخرجوه, which is a translation of και λαβοντες αυτον (*and they took him*). A morphological segmentation results in و + اخرجـ + و + ه which roughly means *and* + *took* + *they* + *him*. The third morpheme indicates that the subject is the 3rd person plural. Since Greek verbs are also inflected with person/number, the third segment (+و+) does not need to be treated separately. Thus, the segments for alignment should be و+اخرجو+ه for the purpose of the alignment.

Additional considerations are related to orthography. Diacritics are marks of various types written above and below the letters of a word; they affect the meaning and pronunciation of the word they compose. For example ضَرَبَ is pronounced *daraba* while ضُرِبَ is pronounced *duriba*; the first is a verb in the active form that means *he hit* while the second is in the passive form and means *he was hit*. Native speakers and expert readers can often do without most diacritics and are able to disambiguate meanings based on context and common usage.

3. Related Works

The works related to this project consist of interlinears, lexicons, concordances, and Greek NT databases that contain some of the data often included in interlinears (e.g. lemmas, glosses, parsing information).

Greek interlinears have existed in the English language as early as 1896 (Estienne, 1896) and have recently been partially developed in Chinese (Lee et al., 2012). Aside from interlinears, Strong's Exhaustive Concordance of the Bible (Strong, 1894) introduced Strong's numbers, which were numbers assigned to the lemmas. These numbers still exist, albeit in public domain works and translations based on earlier versions of the Greek NT.

In the Arabic language, there are two major reference works related to NT Greek. In 1979, Ghassan Khalaf published a manually compiled lexicon and concordance (1979) of Greek words in the NT based on the Smith-Van Dyck Arabic (SVD) translation of the Bible. The work consisted of Greek lemmas/numbers (his own numbering), a list of glosses compiled from the SVD, and concordance entries consisting of the book/chapter/verse reference along with a segment of the SVD containing the translated word in its context. This work covered the most frequent words of the NT, and, thus, understandably, this manually compiled work did not have complete coverage of prominent words and did not account for all occurrences of the article, demonstratives, pronouns, and prepositions.

In 2003, a Greek-Arabic interlinear entitled *New Testament: Interlinear Greek-Arabic* (Feghali et al., 2003) (NTGAI) was published in print form; it was based on the nearly identical (Porter, 2013) NA27 (Nestle et al., 1993) and UBS4 (Aland et al., 1993 2006) Greek texts and featured the Ecumenical Version of the Bible in Arabic (Feghali, 1992) (also known as Good News Arabic Bible). Each page contained an alignment of the Arabic morphemes or blank spaces below the Greek text and the entire Arabic text displayed on a side panel. As discussed in the previous section, this is the result of the segmentation of Arabic words that is necessary for Greek-Arabic alignment.

With regard to text-based databases of the Greek NT, the MorphGNT project (Tauber and Sandborg-Petersen, 2012 2014) is highly relevant. This project has produced two databases of the entire Greek NT, one is based on Tischendorf's 8th edition and the other is based on that of the Society of Biblical Literature (SBLGNT). These databases consist of a word-per-line format that contains the standard data of interlinears previously discussed and address issues of Greek orthography and normalization.

[1] According to Porter, the sources of the Greek NT have 80%-90% agreement and according to F.F. Bruce (quoted by Henry (1999)) none of the differences affect major Christian doctrines.

[2] Textus Receptus refers to a set of standard Greek NT texts of the 16th and 17th centuries (Freedman et al., 2000).

The works covered in this section, when viewed together and in sequence, make the creation of digital Greek-Arabic interlinears a natural and necessary next step.

4. Interlinear Process

What follows is a description of the process identified for creating a two-way aligned Greek-Arabic NT database. The process consists of four steps. Each step is described in terms of the data used, how it was used or modified, and what procedures and tools were used to modify it. The first two steps (Sections 4.1 and 4.2) consist of the preparation of the of the Greek NT data and the Arabic NT data, respectively. The third step (Section 4.3) consists of aligning the two texts. The fourth step (Section 4.4) consists of rendering the results in a readable format. At the time of writing, the first two steps have been completed and the third and fourth steps have been completed for I John 1–2.

4.1. Step 1: Preparation of Greek Texts

Step 1 consists of preparing the Greek NT data, which was located at the *Greek New Testament Texts* site (Robinson, 2013). The texts of interest were three prominent editions of the TR: Stephanus, Elzevir, and Scrivener[3]. These three combine to serve as a good basis for constructing the Mill's text (more about this in Section 5).

The text files of these TR editions were encoded in modified-betacode, an ASCII encoding of the Greek letters. These files were converted to Unicode with a *bash* script that called the *tr* command to map the ASCII characters to Unicode. Each verse began with a chapter/verse reference (e.g., 4:10), but could span many lines. A *Python* script was written to convert the files to a verse-per-line format (VPL). The Stephanus TR contained variant readings (i.e., different spelling of a word or additional words) taken from Scrivener's edition of the TR; thus, a *Python* script was also written to split the file into two texts according to the variant readings. The Elzevir text did not contain variants and was prepared separately.

For every word, the files contained a Strong's number and a parsing code. Some words had a second Strong's number that encoded parsing information; this was eliminated because it was redundant and unintelligible by itself. The VPL files were converted to a word-per-line format (WPL) similar to that of the MorphGNT databases. The fields consisted of a numerical verse/word reference (a 6-digit number equally divided for book, chapter, and verse number), the Greek word, the Strong's number, and parsing codes. A final *Python* script added individual word numbers to the references and a single letter suffix for alignment purposes; this will be explained in Section 4.3). Figure 3 illustrates the flow of the file preparation process, including the main tools used.

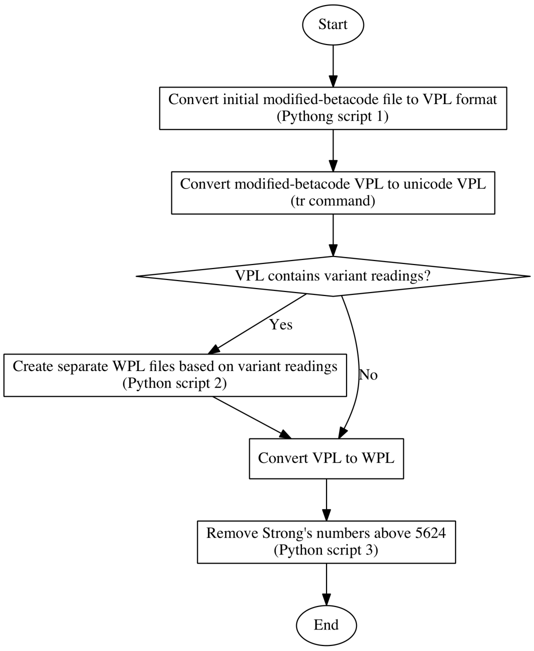

Start
Convert initial modified-betacode file to VPL format (Pythong script 1)
Convert modified-betacode VPL to unicode VPL (tr command)
VPL contains variant readings?
Yes
No
Create separate WPL files based on variant readings (Python script 2)
Convert VPL to WPL
Remove Strong's numbers above 5624 (Python script 3)
End

Figure 3: Flow of process for convert Greek text files to Unicode VPL format.

4.2. Step 2: Preparation of Arabic Text

Step 2 consists of preparing a WPL format of the Arabic text. However, the process was not as straight forward as with the Greek, as issues of Arabic morphology and orthography require further processing.

First, the Arabic NT text was tokenized and a list of unique tokens was created using the NLTK (Bird, 2006). This list was segmented using the Buckwalter Morphological Analyzer (Buckwalter, 2002) which generated one or more morphological analysis solutions for each token. The solutions contained transliterated Arabic text using Buckwalter's own transliteration (2002 2003), thus, Andy Robert's *buckwalter2unicode* script (2004) was modified to render the Arabic portions of the output in Arabic Unicode. A *Perl* script was then written to flatten the multi-line Buckwalter solutions file into a CSV file that contained a single record for each solution; the script also assigned a unique number to each token as a key for later reference. To eliminate multiple solutions, a file was created containing unique four-tuples consisting of token with diacritics, token number, token without diacritics, and segmentation solution (without diacritics).

Additional processing was applied to de-segment morphemes that did not correspond to individual Greek words (as discussed in Section 2). These included the following verbal suffixes: ـتم, ـيا, and ـتا, the plural marker ـات, and ambiguous suffixes that serve both functions ـون, ـين, and

[3] The Scrivener text did not exist at the time of the SVD translation, rather it is a Greek text that was reconstructed to reflect the original text of the King James Version, which used texts that were available at the time of translation.

ـان. The feminine marker ـة was also de-segmented. This file was generated using the nix-based text processing tools *sed*, *awk*, *sort*, and *uniq*.

A *Python* script was developed to aid in eliminating multiple solutions for a single word and to ensure that every word had a complete solution. The process was as follows: (1) append missing segments to incomplete solutions with the missing letters from the original word; (2) for words with multiple solutions, keep the first solution whose segments combine to form the original word, otherwise, simply keep the first solution; (3) indicate if the concatenations of the tokens in the segmentation solution equaled the original word. The script took into consideration orthographic changes such as لـ+الـ contracting to للـ. Other transformations covered 20 tokens and, thus, were fixed manually. Table 1 contains the list of all the changes that have been detected and the means of verification for their segmentation solutions. Automation was employed for cases that were not considered over-generalizations.

Method of Verification	Segmented Text	Intermediate Segmentation or Form	Final Form Form
Automatic	و+لـ+الـ	و+لـ+لـ	وللـ
Automatic	فـ+لـ+الـ	فـ+لـ+لـ	فللـ
Automatic	ة+	ت+	ت
Manual	لـ+الـ	لـ+لـ	لّـ
Manual	نـ+نـ	ننـ	نّـ
Manual	نـ+مـ	ـنمـ	ـمّـ
Manual	يـ+يـ	ـيي	يّـ

Table 1: Orthographic changes considered while verifying word segmentation solutions.

Solutions whose combined segments did not equal the original word and did not contain the changes listed in Table 1 were manually corrected. Out of 24,550 tokens, there remained 1,274 tokens that had no solutions; 679 of them consisted of a single morpheme; 595 required manual segmentation. The manual processing was done in the *vi* editor in which regular-expression based find/replace is rather fluid; this process simply consisted of segmenting the words by matching prefixes and suffixes and confirming their replacement if it resulted in a correct solution.

This file cannot be assumed to be fully correct as ambiguities may arise that show that the wrong solution was chosen or that the assumption of a single solution was false. Simply verifying that all segmented solutions can form the original word allows the interlinear creation process to move into the alignment phase. Errors that are discovered can be corrected in the database of segmented solutions. Correct solutions would be available for the remainder of the alignment process. This process will be discussed in Section 5.

Two limited license Arabic NT texts[4]. were converted to WPL format, cross checked, and corrected. The WPL files contained two columns: the 6-digit numeric verse reference and a token with diacritics, including punctuation. One file was prepared for the alignment phase. Similar to the Greek text, the verse references were augmented with additional word identifiers for the purpose of identifying and sorting the individual morphemes that resulted from the word segmentation.

An additional *Python* script was written to create the final file for alignment. The fields of the file were the following: (1) the augmented reference (verse reference with word number and single-letter suffix), (2) token with diacritics and punctuation, (3) non-punctuated token with diacritics, (4) token without diacritics, (5) morpheme, (6) token with diacritics (no punctuation) in Buckwalter encoding, (7) non-diacritics token in Buckwalter encoding, (8) morpheme in Buckwalter encoding, and (9) word list identifier. Although there is redundancy in format (columns 2, 3, 4, 6,and 7), this does pay off in terms of simplicity and in line-based processing where it gives sufficient context for the word. Such redundancy also exists in the SBLGNT version of the MorphGNT database. In addition, Buckwalter encoding allows for convenient search of the files from the **nix* command line and in editors that do not display Arabic well. The PyArabic library (Zerrouki, 2010) was used by the *Python* scripts to remove diacritics and normalize the Arabic text. Figure 4 illustrates the flow of the entire conversion process, including the main tools used.

4.3. Step 3: Alignment of Texts

The alignment process consisted of aligning the Greek and Arabic files that resulted from the first two steps in two side-by-side split panes whose scrolling was synchronized. The tool used was the *vi* editor. As mentioned earlier, each word was given a number and a letter suffix; this was to ensure that the order of the words could be recovered when the two files were combined to generate either an interlinear or a reverse interlinear. The alignment, then, consisted of (1) moving lines in the Greek file or the Arabic file to align equivalents and (2) inserting gaps in either file to ensure a proper alignment.

The final letter in the verse/word reference was edited to reflect the new order introduced. The files were initially generated with word reference suffixes in the middle of the alphabet (*l*, *m*, *n*, and *o*). Gaps before a word were given a suffix of *a*; gaps after a word were given a suffix of *z*; gaps in the middle required the letters to be changed to reflect the correct order.

The gaps were required for instances where Greek words were not translated (e.g., the article[5]) as well as instances

[4] An unrestricted version of the SVD is available from the Multilingual Bible Parallel Corpus (Christodoulopoulos, 2014)

[5] There is no indefinite article in Biblical Greek and definiteness can be marked without it. See (Wallace, 1996) for further

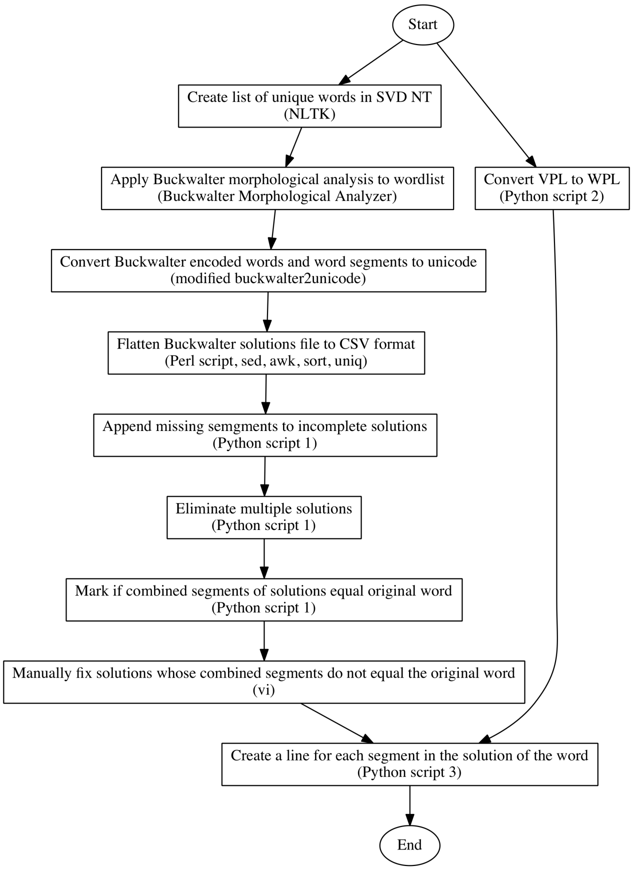

Start
Create list of unique words in SVD NT (NLTK)
Apply Buckwalter morphological analysis to wordlist (Buckwalter Morphological Analyzer)
Convert VPL to WPL (Python script 2)
Convert Buckwalter encoded words and word segments to unicode (modified buckwalter2unicode)
Flatten Buckwalter solutions file to CSV format (Perl script, sed, awk, sort, uniq)
Append missing semgments to incomplete solutions (Python script 1)
Eliminate multiple solutions (Python script 1)
Mark if combined segments of solutions equal original word (Python script 1)
Manually fix solutions whose combined segments do not equal the original word (vi)
Create a line for each segment in the solution of the word (Python script 3)
End

Figure 4: Flow of process for convert Arabic text files to VPL format.

Type	Description	Symbol
No-Equivalent	No equivalent for the token on the same line in the other file	-
Pre-Equivalent	Following word corresponds to two or more tokens in the other file	\
Post-Equivalent	Previous word corresponds to two or more tokens in the other file	/

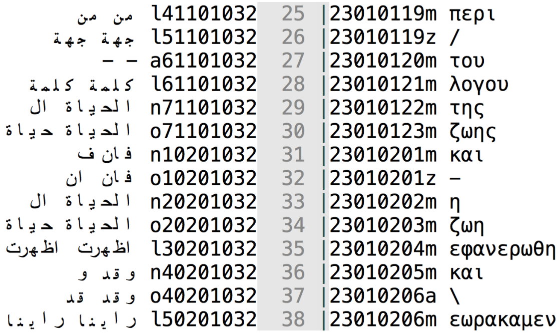
Table 2: Types of gaps identified in alignment process.

where Arabic words/morphemes were added that had no equivalent in the Greek text (e.g., the definite article or words to make for a more readable translation). Gaps were inserted before, inside, or after a word and were given a reference that preserved the order of the file. Three types of gaps were identified, they are listed in Table 2. Figure 5 contains an example of each: a *no-equivalent* is found on line 32, *pre-equivalent* on line 37, and *post-equivalent* on line 26.

In cases involving gaps, precedence was given to aligning the two tokens from each language that matched more literally or in essence while setting apart what matched less. For example, υμιν is the 1st person dative plural pronoun which in English is translated *to you*; it was translated to اليكم (اليـ+ـكم), whose two segments form a literal equivalent to the English translation. In this case, the pronoun υμιν was aligned with the second segment ـكم which is recognized in Arabic as pronoun. The preposition الى [6] was matched with a pre-equivalent gap since it was a preposition.

Another example is εωρακαμεν, a verb in the perfect tense (1st person plural) which is translated *we have seen*. In the literal Arabic translation قد راينا, the first word is a particle that conveys the meaning of *have* and the second word is segmented into رايـ+ـنا (roughly translated *saw*+*we*), thus, the قد was matched with a pre-equivalent gap while راينا was matched with εωρακαμεν. In instances where there is no basis for judging near equivalence, the second word was treated as a *post-equivalent* gap. This may be indicated with a different symbol in the future in order to differentiate it from *post-equivalent* gaps.

In addition, there were cases where a gap in the Greek text was added for an Arabic word with a segmentation solution. For example, there was an instance where به (بـ+ـه) was added to the Arabic translation for clarity. In such cases, the segments were joined into a single word/line entry since the purpose of segmenting Arabic words was to correctly map the segments to corresponding Greek words and not for the sake of Arabic morphological analysis.



```
      من من l41101032  25 |23010119m περι
    جهة جهة l51101032  26 |23010119z /
        - - a61101032  27 |23010120m του
  كلمة كلمة l61101032  28 |23010121m λογου
  الحياة ال n71101032  29 |23010122m της
الحياة حياة o71101032  30 |23010123m ζωης
      فان ف n10201032  31 |23010201m και
     فان ان o10201032  32 |23010201z -
  الحياة ال n20201032  33 |23010202m η
الحياة حياة o20201032  34 |23010203m ζωη
اظهرت اظهرت l30201032  35 |23010204m εφανερωθη
      وقد و n40201032  36 |23010205m και
     وقد قد o40201032  37 |23010206a \
راينا راينا l50201032  38 |23010206m εωρακαμεν
```

Figure 5: Sample of alignment of Greek and Arabic text.

4.4. Step 4: Generating a Displayable Format

The final step was to generate a displayable format for the interlinear and reverse interlinear. This involved joining the two aligned files into a single file, which was accomplished with the *paste* command. In order to preserve the Greek word order for the interlinear, the file was sorted by the column containing the Greek verse/word references. Likewise, to generate the reverse interlinear, the Arabic order was acquired by sorting by the column containing

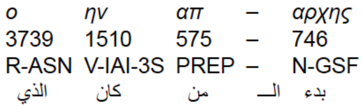
discussion.

[6] الى takes on the form اليـ takes on when combined with ـكم

the Arabic verse/morpheme references.

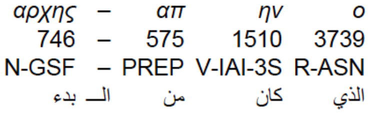
The displayable format that was generated was a prototype in PDF format. It consisted of a PDF file that was generated with the *Latex* text formatting system. The *Latex* files used the *expex* and *polyglossia* packages, which format interlinear text and display multilingual text, respectively. The *Latex* code was generated with a *bash* script that used a combination of nix-based text processing tools (*cat*, *sort*, *sed*, *grep*, *cut*, *tr*. Figures 6 and 7 show a sample for each file.

The aligned text consisted of the Greek text, Strong's number, the parsing codes, and the Arabic morphemes. The Arabic was placed on the bottom line in the reverse interlinear due to a problem faced with displaying Arabic text on the first line; ideally, it should be on top. *No-equivalent*, *pre-equivalent*, and *post-equivalent* gaps were represented as dashes (–), either a forward-pointing arrows (→) or a back-pointing arrow (←). The direction of the arrows changed according to which language had its word order preserved. Section 6 will have more to say about other displayable formats.



ο	*ην*	*απ*	–	*αρχης*
3739	1510	575	–	746
R-ASN	V-IAI-3S	PREP	–	N-GSF
الذي	كان	من	الــ	بدء

Figure 6: Beginning of I John 1 in interlinear format.



αρχης	–	*απ*	*ην*	*ο*
746	–	575	1510	3739
N-GSF	–	PREP	V-IAI-3S	R-ASN
بدء	الــ	من	كان	الذي

Figure 7: Beginning of I John 1 in reverse interlinear format.

Although the displayable format is still in the prototype phase, it is still possible to compare the project's output (GrArNT) with the work of Feghali et. al (NTGAI). The comparison in Table 3 is in terms of format and available data. The inclusion of the Greek lemma is feasible and its inclusion will also be discussed in section 6..

Feature	NTGAI	GrArNT
Digital Text Format	No	Yes
Interlinear	Yes	Yes
Reverse Interlinear	No	Yes
Strong's Number	No	Yes
Lemma	No	Feasible
Parsing Information	No	Yes

Table 3: Comparison between NTGAI and GrArNT.

5. Research Findings

Although the project is in the early phases of the alignment process, there have been many interesting findings pertaining to (1) the source text of the SVD NT, (2) the prevalence of Arabic words with segments, (3) issues pertaining to the correctness of segmentation solutions, and (4) the number and kinds of edits that took place during the alignment.

Mill's TR was based on Stephanus' 1550 edition and according to Bagster and Sons' printing of Mill's Greek text (Bagster, 1851), it contains "one correction and a few unintentional changes" and it "differs from the Elzevir text in comparatively few places." Thus, it was decided to reconstruct Mill's TR using the three available TR editions during the alignment process while checking the text against the Mill TR (Bagster, 1851; Mill, 1707) . The need for all three texts (Stephanus, Elzevir, and Scrivener) was confirmed with a tree-way comparison that showed that there were instances where the SVD favored a reading that existed in only one of the three texts (e.g., Luke 10:22, John 12:17, and Luke 1:35, respectively).

The alignment process of I John 1-2 further verified the usefulness of all three texts when a variant readings was encountered in verses 4. In verse 4, Stephanus has ημων (*your*), while Elzevir and Scrivener have υμιν (*our*) which is also what Mill's text contained[7]. Such findings allow this project to contribute knowledge regarding the textual roots of the SVD NT and to account for the differences between Mill's and Stephanus' texts.

Another finding related to Mill's TR and the SVD NT translation is that an instance was found where the SVD translators chose a reading not found in Mill's TR. In I John 2:23 Mill, Sephanus, and Elzevir read πας ο αρνουμενος τον υιον ουδε τον πατερα εχει, but SVD adds ο ομολογων τον υιον και τον πατερα εχει, a reading included in modern texts and translations (Metzger, 1994)[8].

A second finding is related to the number of multi-segment Arabic words vs. those consisting of a single morpheme. The total number of tokens in the text was 606, which consisted of 329 unique words. 42% of the word occurrences were multi-segment words; 47% of the words in the unique word list were multi-segmented. Although similar figures could be computed for the entire SVD NT, aligned and confirmed text would make for more reliable figures.

A third finding is related to the correctness of the segmentation solutions. In I John 1:2, the solution for راينا was ـنا+ رايـ (*our opinion*) instead of رايـنا (*we saw*). This occurred because the Buckwalter morphological analyzer dis-

[7] The difference results in the verse 4 reading "And these things we write, in order that [your] joy may be complete." instead of "... so that our joy may be complete." (LEB(Harris et al., 2012), first translation modified by author)

[8] The full verse reads "Everyone who denies the Son does not have the Father either; *the one who confesses the Son has the Father also*" (LEB, the italicized text is the additional text).

regarded diacritics. In addition, the suffix ـنا could indicate the objective case when it suffixed a verb.

As a result, the database of solutions was searched for these ambiguities and the segmentation solutions were fixed manually. Out of 480 distinct Arabic tokens aligned, this was the only case of a solution that needed correction. However, this is only the beginning of the alignment and more corrections are expected.

Type	Greek	Arabic
No-Equivalent	105	54
Pre-Equivalent	33	0
Post-Equivalent	11	9

Table 4: Number of gaps added to each text during the alignment of I John 1 & 2.

The last set of findings is related to how the alignment was affected by the translation. The Arabic text for these two chapters consisted of 883 total tokens to align (including segments of words), while the Greek text consisted of 796. During the alignment, 2% of Arabic morphemes occurrences had their order changed and 11% of Greek word occurrences had their order changed. 63 gaps were inserted into the Greek text while 149 gaps were introduced into the Arabic text. Table 4 contains a breakdown of the number of gaps for each type in their respective texts. The most frequent gaps for both the Greek and Arabic texts were of the *no-equivalent* type; they were for instances where the article (ο, الـ) was found in one language without a counterpart in the other. Gaps consisted of 42% of the changes to the Greek text and 88% of the changes to the Arabic text.

This body of findings is expected to grow as more of the NT is aligned. By the end of the alignment, the Mill TR will be fully constructed and all its variations from the Stephanus TR will be known. SVD variations from Mill's text will also be known. Although the text for the alignment was chosen due to its familiarity and usefulness for instruction in Greek, it contained a representative sample of the types of findings available regarding the source text and nature of the Arabic translation. The findings regarding the prominence of *pre-equivalent* gaps will inform the development of a safe algorithm to align the texts.

6. Future Possibilities

There are several future possibilities related to enhancing the current database in terms of linguistic and textual information as well as extending and replicating its usefulness.

The aligned database could be augmented with information that enhances readability such as Greek lemmas, instead of only Strong's numbers, and transliterations. Arabic glosses and parsing information would also be useful for translational insight. It is also necessary and feasible to add accents to the current Greek text, which does not have them. Finally, clause-level alignment of the texts should be investigated.

Second, the interlinear can be published in more formats. Converters can be written to generate XML-based formats such as TEI (TEI, 2014) and OSIS (OSIS, 2014). OSIS, an XML standard for the encoding of the Bible, is an especially desired target because there are OSIS converters that create modules for Bible software (crosswire.org, 2014; Smith, 2006). The current *Latex* generated PDF format could also be made more compact. Finally, a user configurable web interface would allow a dynamic and interactive reading experience.

Third, it would be helpful to replicate this process for other Arabic translation of the NT, both ancient and modern. It may also be possible to bootstrap such efforts using the files that were already generated.

Fourth, it is possible to carry out corpus-based studies and translation evaluations. Some important issues to investigate would be (1) features of Arabic not found in Greek such as the dual, (2) the translation of prepositions, and (3) the article in the Greek text vs. in the Arabic text.In addition, as early as 1956 (Thompson, 1956) and as recently as 2009 (Khalaf, 2009), there have been published discussions regarding the revision of the SVD translation. The electronic search of the aligned texts would be helpful in surveying the translation for possible changes and in ensuring the consideration of all instances of words being investigated. Finally, other Arabic translations, once aligned to their respective text, can be compared with other Arabic translations in terms of the source text and translation; this would help trace sources of differences between them.

Fifth, Mill's TR and other similar texts do not reflect the latest Greek manuscripts and texts. Thus, it would be beneficial to indicate readings not found in Mill's TR from more recent Greek NT editions.

Finally, automation can begin to be introduced into the alignment process based on the gaps related to the article in Greek and the definite article in the Arabic as well as other patterns that are certain. Furthermore, statistically-based and lookup-based alignment methods could be explored with the aim of achieving a more generic automated solution that would help bootstrap the process of creating Greek interlinears for NT translations in other languages. In addition, this interlinear process could be extended to other works in Greek such as the Septuagint[9] and ancient Greek works that have been translated into Arabic.

[9] The Septuagint or LXX is the Greek translation of the Old Testament used before and throughout during the New Testament era and in the early church

These possibilities would enrich the study of the New Testament and its Arabic translations by connecting them thoroughly to their source texts, and by allowing corpus-based Arabic NT investigations. Generalized solutions for other languages would also facilitate the study of NT translations for which no interlinears exist.

7. Conclusion

Having digital Greek-Arabic interlinears is of high benefit to all investigations into the New Testament. This paper has shown how the proper resources, tools, and linguistic considerations can result in a two-way aligned database from which interlinears of both kinds, regular and reverse, can be generated. Thus, it helps set a basis for how other Arabic translations of the NT can have corresponding interlinears. The alignment project has implications for textual studies as it facilitates reconstructing the original Greek text of the SVD NT in light of the discovery that it has readings not found in Mill's TR. This project is also able to yield many benefits in the future by facilitating corpus-based investigations of the text that go beyond manual analysis and yield a more comprehensive view of the relationship between source text and translation. Its most immediate benefits are for Biblical Greek pedagogy in the Arab-speaking world as well as for research and discussion of the NT based on Arabic translations.

8. Acknowledgements

I would like to thank my family, church, and congregation for their spiritual support and encouragement, AUB for the opportunities and support it has given this project, and N.E.S.T. for its warm welcome and help in research.

9. References

Aland, K., Black, M., Martini, C. M., Metzger, B. M., Robinson, M., and Wikgren, A. (1993; 2006). *The Greek New Testament*. Deutsche Bibelgesellschaft, 4th edition.

Bagster, S. (1851). *He Kaine Diatheke. The New Testament, the 'Received Text' with selected various readings from Griesbach, Scholz, Lachmann and Tischendorf and references to parallel passages*. S. Bagster & Sons.

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Buckwalter, T. (2002). Buckwalter arabic morphological analyzer version 1.0. `http://catalog.ldc.upenn.edu/LDC2002L49`. Accessed: 2014-02-25.

Christodoulopoulos, C. (2014). Multilingual bible parallel corpus. `http://homepages.inf.ed.ac.uk/s0787820/bible/`. Accessed: 2014-02-25.

crosswire.org. (2014). Crosswire bible society. `http://crosswire.org`. Accessed: 2014-02-25.

Estienne, R. (1896). *The Englishman's Greek New Testament*. London, 3rd edition.

Feghali, B., Aoukar, A., Khouri, N. A., and Fakhri, Y. (2003). *New Testament: Greek-Arabic Interlinear*. Antonine University.

Feghali, B., editor. (1992). *Good News Arabic*. Bible Society.

Freedman, D. N., Myers, A. C., and Beck, A. B. (2000). *Eerdmans dictionary of the Bible*. W.B. Eerdmans.

Hall, I. H. (1885). The arabic bible of drs. eli smith and cornelius va van dyck. *Journal of the American Oriental Society*, 11:276–286.

Harris, W. Hall, I., Ritzema, E., Brannan, R., Mangum, D., Dunham, J., Reimer, J. A., and Wierenga, M. (2012). *The Lexham English Bible*. Lexham Press.

Khalaf, G. (2009). *Shedding Light on Boustani*. Bible Society.

Lee, J., Wong, S. S. M., Tang, P. K., and Webster, J. (2012). A greek-chinese interlinear of the new testament gospels. In *LREC'2012 Workshop: LRE-Rel – Language Resource and Evaluation for Religious Texts*, pages 42–48, Istanbul, Turkey, May.

Metzger, B. M. (1994). *A Textual Commentary on the Greek New Testament*. United Bible Societies, 2nd edition.

Mill, J. (1707). John mill – novum testamentum. `http://www.csntm.org/printedbook/viewbook/JohnMillNovumTestamentum1707`. Accessed: 2014-02-25.

Nestle, E., Nestle, E., Aland, B., Aland, K., Karavidopoulos, J., Martini, C. M., and Metzger, B. M. (1993). *The Greek New Testament*. Deutsche Bibelgesellschaft, 27th edition.

OSIS. (2014). Osis. `http://www.bibletechnologies.net`. Accessed: 2014-02-26.

Porter, S. E., (2013). *The Text of the New Testament*, pages 50–76. Acadia Studies in Bible and Theology. Baker Academic.

Robinson, M. (2013). Greek new testament texts. `https://sites.google.com/a/wmail.fi/greeknt/home/greeknt`. Accessed: 2014-02-24.

Smith, E., (1900). *Report of Rev. Eli Smith, D.D. on the translation of the Scriptures, April, 1834*, page 10. American Presbyterian Mission Press. Courtesy of N.E.S.T. Special Collections.

Smith, D. (2006). Project kjv 2006. `http://crosswire.org/~dmsmith/kjv2006/`. Accessed: 2014-02-25.

Strong, J., editor. (1894). *Strong's Exhaustive Concordance of the Bible*. Hendrickson Publishers.

(1865). *Arabic Bible (Smith-Van Dyck)*. Bible Society.

Tauber, J. and Sandborg-Petersen, U. (2012–2014). Morphgnt. `http://morphgnt.org`. Accessed: 2014-02-25.

TEI. (2014). Text encoding initiative. `http://www.tei-c.org`. Accessed: 2014-02-26.

Thompson, J. (1956). *The Major Arabic Bibles: Their Origin and Nature*. American Bible Society.

Wallace, D. B. (1996). *Greek Grammar beyond the Basics: An Exegetical Syntax of the New Testament*. Zondervan.

Zerrouki, T. (2010). Pyarabic 0.2. `https://pypi.python.org/pypi/PyArabic/0.2`. Accessed: 2014-02-24.

A Proposed Model for Quranic Arabic WordNet

Manal AlMaayah[1], Majdi Sawalha[2], Mohammad A. M. Abushariah[3]
Computer Information Systems Department, King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan.
E-mail: [1]manalmaayah@ymail.com, [2]sawalha.majdi@gmail.com, [3]m.abushariah@ju.edu.jo

Abstract

Most recent Arabic language computing research focuses on modern standard Arabic, but the classical Arabic of the Qur'an has been relatively unexplored, despite the importance of the Qur'an to Islam worldwide which can be used by both scholars and learners.This research work proposes to develop a WordNet for Qur'an by building semantic connections between words in order to achieve a better understanding of the meanings of the Qur'anic words using traditional Arabic dictionaries and a Qur'an ontology. The Qur'an corpus will be used as text and Boundary Annotated Qur'an Corpus (Brierley et al, 2012) will be used to explore the root and Part-of-Speech for each word and the word by word English translations. Traditional Arabic dictionaries will be used to find the Arabic meaning and derived words for each root in the Qur'anic Corpus. Then, these words and their meanings (Arabic, English) will be connected together through semantic relations. The achieved Qur'anic WordNet will provide an integrated semantic Qur'anic dictionary for the Arabic and English versions of the Qur'an.

Keywords: Qur'an WordNet, Qur'an ontology, Arabic dictionaries.

1. Introduction

The Holy Qur'an is the central religious text of Islam. The Qur'an is considered to be an excellent gold standard text that is essential for developing, modeling and evaluating Arabic NLP tools. The Qur'an as a corpus and is made up of 77,430 words. It is divided into 114 chapters which consist of 6,243 verses. The Qur'anic WordNet services anyone who seeks to expand his knowledge of Qur'anic Arabic vocabulary and increases understanding of the Qur'an and of Islam. In Qur'an, we find many words that are conceptually synonyms but if we investigate their dictionary meanings, then differences will surface. For example:أحمَد*ahmad* (SAW), المزَّمِّل*al-muzzammil*, المدَّثِّر*al-muddattir*, يس *yāsīn*, الرَّسول*ar-ras˙l*are synonymous of Muhammad (SAW). Another example is سَبِيلِ *sabīl* and وَجْهِ *wağh* are synonyms of (the way that they spend their wealth). Table 1 illustrates this example.

Table 1: Examples of سَبِيلِ sabīl and وَجْهِ wağh in different verses.

Example 1: Chapter 2, verse 272
وَمَا تُنْفِقُونَ إِلَّا ابْتِغَاءَ وَجْهِ اللَّهِ
and you do not spend except seeking the countenance of Allah
Example 2: Chapter 2, verse 261
مَثَلُ الَّذِينَ يُنْفِقُونَ أَمْوَالَهُمْ فِي سَبِيلِ اللَّهِ كَمَثَلِ حَبَّةٍ أَنْبَتَتْ سَبْعَ سَنَابِلَ
The example of those who spend their wealth in the way of Allah is like a seed [of grain] which grows seven spikes

Therefore, we will get a better understanding of the meanings of the Qur'an by modeling a Qur'anic WordNet and developing a computational linguistic theory for Arabic using new technologies of NLP, traditional Arabic linguistic theory and classical Arabic dictionaries. We can utilize Qur'anic WordNet for machine learning tasks, such as Word Sense Disambiguation (WSD) where the word may have many meanings and it becomes therefore crucial to distinct the different senses. For example, the word وجه *wağh* has three senses as illustrated in Table (2). To decide the returned sense in a current context, the Qur'anic WordNet is essential.

Table 2: Examples of different senses of the word وجه *wağh*

لِلَّهِ وَمَنِ اتَّبَعَنِ	وَجْهِيَ	فَإِنْ حَاجُّوكَ فَقُلْ أَسْلَمْتُ
Submitted myself to Allah		
النَّهَارِ	وَجْهَ	آمِنُوا بِالَّذِي أُنزِلَ عَلَى الَّذِينَ آمَنُوا
At the beginning of the day		
	وَجْهِهَا	ذَٰلِكَ أَدْنَىٰ أَن يَأْتُوا بِالشَّهَادَةِ عَلَىٰ
It's true form		

This paper is structured as follows: section 2 a brief overview of the Qur'anic WordNet, section 3 related work, section 4 methodology, section 5 conclusion and future work.

2. A Brief Overview of the Qur'anic WordNet

The Qur'anic WordNet is a multidisciplinary project that contributes to Information Technology including; Computational Linguistics and Language Engineering, Information Extraction, Text Analytics, Text Data Mining, and Machine Learning, and to other disciplines such as Islamic Studies, Linguistics, Arabic Linguistics and Lexicography.

Qur'anic WordNet will make use of Arabic WordNet (Elkateb et al, 2006), Qur'an Ontology and classical

Arabic dictionaries. We will provide a literature investigation on WordNets, Arabic WordNet and Qur'an WordNet and ontologies. Then we will design a method for measuring semantic similarity between words included in the Qur'anic WordNet.

3. Related Work

Several research initiatives were directed towards building a WordNet for various foreign languages. WordNet was first developed for English in 1980s at the Cognitive Science Laboratory of Princeton University, hence is known as Princeton WordNet. This is a large-scale lexical database that was manually constructed (Miller,1995), (Miller and Fellbaum, 2007), (Fellbaum and Vossen, 2012).

George A. Miller (1995) showed the importance of WordNet for the English language and he outlined the semantic relations that provide more effective combinations of traditional lexicographic information and modern computing. English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms, each representing a lexicalized concept, and semantic relations which link the sets of synonym. He used more than 116.000 pointers between words and word senses to build these relations. The semantic relations that were used in WordNet include (synonymy, antonymy, hyponymy, meronymy, troponomy, entailment), in addition to the senses. The interface of WordNet is easy to use with all word forms. When a word is entered, its syntactic category appears in menu. Once the appropriate syntactic category is selected, semantic relations for that word are displayed.

In 1998, the EuroWordNet of eight European languages was developed and linked to the English WordNet. This is a resource that provides multilingual lexical database. EuroWordNet contributed to variety of fundamental innovations in the design of the WordNet. It defines a set of base concepts and increases the connectivity among synsets (Fellbaum and Vossen, 2012).

In 2000, the idea of the Global WordNet was founded by Fellbaum and Vossen (Fellbaum and Vossen, 2007) and (Vossen and Fellbuam, 2012) for the aim of establishing a WordNet that supports a large number of languages interlinked into a single knowledgebase to facilitate inter communicability. Currently, Global WordNet covers sixty distinct languages, including: Arabic, Bantu, Basque, Chinese, Bulgarian, Estonian, Hebrew, Icelandic, Japanese, Kannada, Korean, Latvian, Nepali, Persian, Romanian, Sanskrit, Tamil, Thai, Turkish, Zulu, etc.

Elkateb and Black (2006) introduced the project for building a lexical resource for Modern Standard Arabic based on the English language's Princeton WordNet (Fellbaum, 1998) and the standard word representation of senses. The tool that was built has a lexicographer's interface and can be linked directly to EuroWordNet (EWN).

Trad and Koroni (2012) used Arabic WordNet Ontology for query expansion. They evaluated how beneficial was it compared to the English WordNet Ontology. Two individual corpora were used, one for Arabic documents and the other for English documents. Scanning and indexing files and documents are made via a multithreaded procedure to maintain the best interactivity and efficiency. The result of the comparison showed that the English ontology was better and more global than the Arabic ontology.

Shoaib (2009) proposed a model that is capable of performing semantic search in the Qur'an. The model exploits WordNet relationships in a relational database model. The implementation of this model has been carried out with the latest technologies and Surah Al-Baqrah (Chapter 2 from the Qur'an) has been taken as a sample text. The model facilitates performing a subject search for Qur'an readers and provides a framework capable of retrieving related verses from the Qur'an whether the query keyword is present in them or not. Semantic search is carried out in two steps. The first step is to identify only one sense of the query keyword using (WSD). The second step is to retrieve all synonyms of the identified sense of the word. The main goal of this work is to improve the semantic search and retrieval over Qur'anic topics.

4. Our proposed model for the Arabic Qur'anic WordNet

The proposed work "Arabic Qur'anic WordNet" will be implemented and evaluated (as illustrated in figure 1) in the following steps:

- **Qur'an text is preprocessed** using (i) tokenization, (ii) elimination of stop words, (iii) stemming and (iv) Part-Of-Speech tagging for each word in the Qur'an corpus.

- **Synsets (synonym sets)** are generated by grouping words of similar meaning and part-of-speech. For example, the words {رأى *rā* , أبصَر *abaṣara*, نظَر *nazara*} that share the sense, "see", are grouped together in a synset.

- **Semantic relations between different synsets are defined**. The semantic relationships that will be included in the Qur'anic WordNet are:

  a. Synonymy is determined. The words that have similar meanings are synonyms. For example, the words {حول *ḥawl* , سنه *sanah*, عام *ʿām*} are synonyms, they all mean "a year".
  b. Antonyms are marked. The words that have opposite meanings such as الحياة *al- ḥayāt* "life", and الموت *al-mawt* "death" are labeled 'antonyms'.
  c. A Glossary is compiled. This is used to store the glosses for every synset. Gloss may contain an

explanation, definition, and example of sentences. For instance, {المطر *al-maṭar*, الغيث *al-ġayṯ*} is a synset that share the sense "rain". However, they are used in different contexts (see Table 3).

d. **Similarity** between concepts is differentiated by connecting synsets that have similar meanings. For example, {خشيه *khāshīnyh*, خوف *khawf*} and {الروع *ar-rawʿu*, الرهب *ar-rahb*} is a synset that share the meaning of "fear" or "fright", see Table 4.

The implementation of the Qur'anic WordNet utilizes the Qur'an corpus as text and the Boundary Annotated Qur'an Corpus (Brierley et al, 2012) for exploring roots, POS tags, and English meaning for each word in the corpus. After that, the Arabic meaning and the derived words for each root are found using classical Arabic dictionaries. Finally, these words and their meanings (Arabic, English) are connected together with semantic relations.

The Qur'anic WordNet will be evaluated using a suitable evaluation technologies, standards, and metrics. We will build a gold standard for evaluating Qur'anic WordNet which at the same time can be used for evaluating Arabic WordNet.

Table 3: Glossary Example

Word	Semantics	Example	Translation
al-maṭar المطر	torment	وَأَمْطَرْنَا عَلَيْهِمْ مَطَرًا فَانْظُرْ كَيْفَ كَانَ عَاقِبَةُ الْمُجْرِمِينَ (الأعراف: 84)	And We rained upon them a rain [of stones]. Then see how was the end of the criminals.
al-ġayṯ الغيث	goodness and grace	وَهُوَ الَّذِي يُنَزِّلُ الْغَيْثَ مِنْ بَعْدِ مَا قَنَطُوا وَيَنْشُرُ رَحْمَتَهُ ۚ وَهُوَ الْوَلِيُّ الْحَمِيدُ (الشورى: 28)	And it is He who sends down the rain after they had despaired and spreads His mercy. And He is the Protector, the Praiseworthy.

Table 4.a: Example of similar meaning{خشيه khāshīnyā, خوف khawf}

Word	Semantics	Example	Translation
خشيَه *khāshīnyā*	as (they) fear	إِذَا فَرِيقٌ مِنْهُمْ يَخْشَوْنَ النَّاسَ كَخَشْيَةِ اللَّهِ (النساء: 77)	at once a party of them feared men as they fear Allah or with [even] greater fear.
خوف *khawf*	fear	فَمَا آمَنَ لِمُوسَىٰ إِلَّا ذُرِّيَّةٌ مِنْ قَوْمِهِ عَلَىٰ خَوْفٍ مِنْ فِرْعَوْنَ وَمَلَئِهِمْ أَن يَفْتِنَهُم (يونس: 83)	But no one believed Moses, except [some] youths among his people, for fear of Pharaoh and his establishment that they would persecute them.

Table 4.b:Example of similar meaning{الروع *ar-rawʿu*, الرهب *ar-rahb*}

Word	Semantics	Example	Translation
الروع ***l-rawʿu***	fright	فَلَمَّا ذَهَبَ عَنْ إِبْرَاهِيمَ الرَّوْعُ وَجَاءَتْهُ الْبُشْرَىٰ يُجَادِلُنَا فِي قَوْمِ لُوطٍ (هود:74)	And when the fright had left Abraham and the good tidings had reached him, he began to argue with Us concerning the people of Lot.
الرهب ***l-rahb***	fear	وَاضْمُمْ إِلَيْكَ جَنَاحَكَ مِنَ الرَّهْبِ (القصص:32)	And draw in your arm close to you [as prevention] from fear

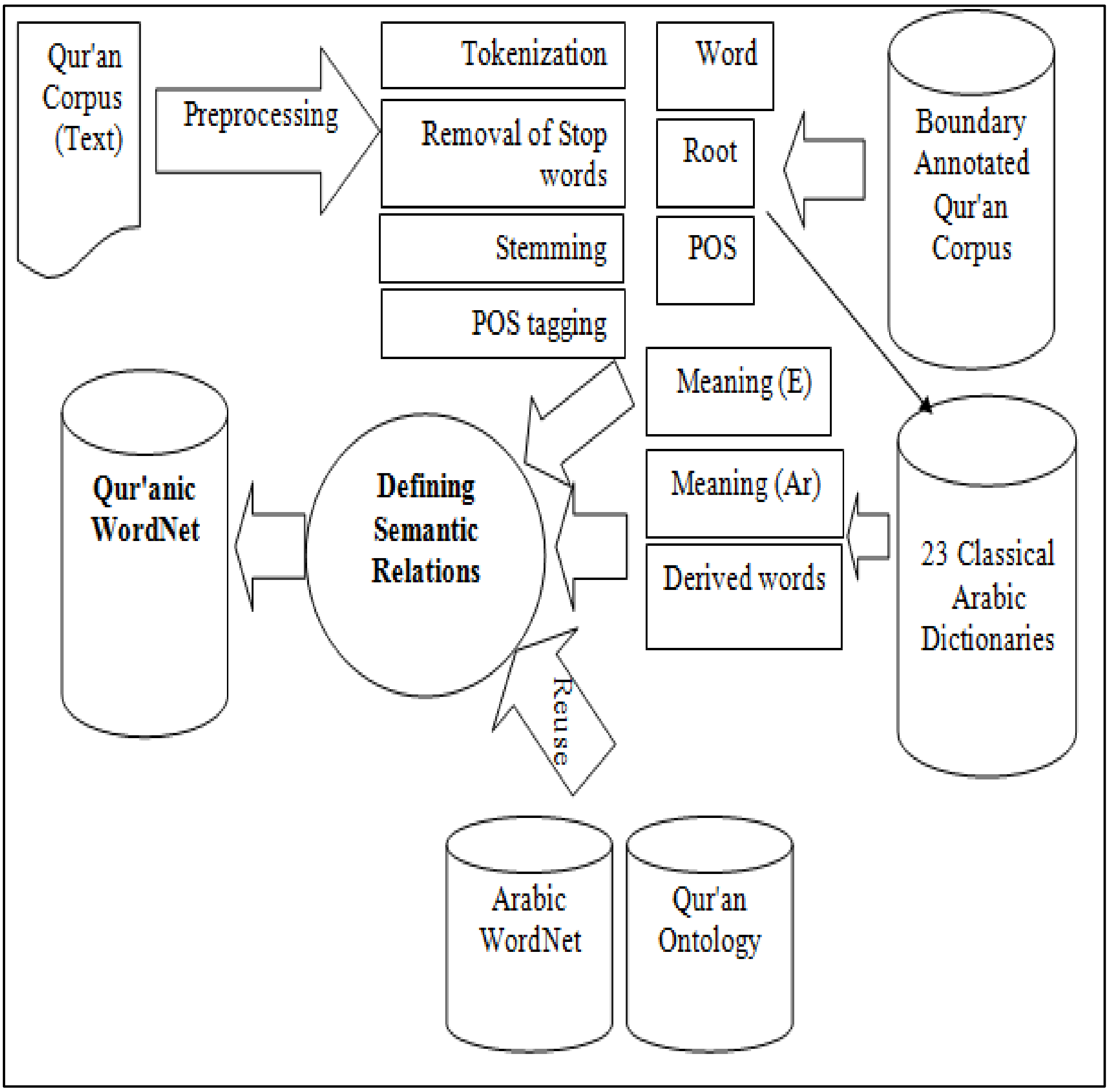


Figure 1: Implementation and evaluations model for building the Qur'anic WordNet

5. Conclusion

This work proposes the Arabic Qur'anic WordNet as a valuable resource designed and built for the study the semantic relations between Arabic concepts in the Qur'an. This work is unprecedented. The proposed model will be implemented on Qur'anic words using Arabic WordNet, classical Arabic dictionaries, and Qur'an ontology. Suitable evaluation methods will be designed, implemented and applied. These shall also be used as a standard for evaluating Arabic WordNet. We will design a methodology for measuring semantic similarity between different words included in the Qur'anic WordNet using edge-counting techniques. After building the Quranic WordNet, we plan to facilitate it for machine learning tasks such as word sense disambiguation.

References

Black, W., Elkateb, S., & Vossen, P. (2006). Introducing the Arabic wordnet project. In *Proceedings of the third International WordNet Conference (GWC-06).*

Brierley, C., Sawalha, M. Atwell E. (2012). open-source boundary-annotated corpus for Arabic speech and language processing. *LREC 2012, Istanbul, Turkey*.

Dukes, K. (2012). The Qur'anic Arabic Corpus. Online. *http://corpus.quran.com* .

Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Building a wordnet for arabic. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006).*

Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

Fellbaum, c. Vossen P. 2012. Challenges for a multilingual WordNet. *Lang Resources & Evaluation*, 46, 313–326.

Fellbaum, C., Vossen, P. (2007). Connecting the universal to the specific. In T. Ishida, S. R. Fussell & P. T. J. M. Vossen (Eds.), Intercultural collaboration: *First international workshop* (Vol. 4568, pp. 1–16). *Lecture Notes in Computer Science*, Springer, New York.

Khan, H., Saqlain, S.M., Shoaib, M., & Sher, M. (2013). Ontology Based Semantic Search in Holy Quran. *International Journal of Future Computer and Communication*. 2, 560-575.

M. Shoaib, M. N. Yasin, U. K. Hikmat, M. I. Saeed, & M. S. H.Khiyal.(2009). Relational Word Net model for semantic search in holy quran. In *Proceedings of International Conference on Emerging Technologies, Pakistan.* pp. 29-34.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*. 38, 39–41.

Miller G. A., Fellbaum C. (2007). WordNet then and now. *Lang Resources & Evaluation.* 41, 209–214.

Muhammad, A. B. (2012). Annotation of conceptual co-reference and text mining the Qur'an. *University of Leeds.*

Rodríguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Antonia Martí, M., Black, W., Elkateb, S., Kirk,J., Pease, A., Vossen, P.,& Fellbaum, C. (2008). Arabic WordNet: Current State and Future Extensions .In *Proceedings of the Fourth International GlobalWordNet Conference-GWC 2008. Szeged, Hungary.*

Sharaf, A., Atwell, E. (2012). QurSim: A corpus for evaluation of relatedness in short texts. *LREC 2012, Istanbul, Turkey.*

Trad, R., Mustafa, H., Koroni, R., & Almaghrabi, A. (2012). Evaluating Arabic WordNet Ontology by expansion of Arabic queries using various retrieval models. In *ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2012 10th International Conference on* (pp. 155-162). IEEE.

Vossen, P. (2002). EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document.

Traditional vs. Chronological Order: Stylistic Distance Analysis in Juz' Amma

Ahmad Alqurneh, Aida Mustapha

Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
ahmad.alqurneh@outlook.com, aida_m@upm.edu.my

Abstract

This paper analyzes the smoothness sequence in different texts in order to investigate whether texts that are close in time (which implies they are in chronological order) are stylistically similar or otherwise. We propose a spatial metaphor of distance with graphical representation to show similarity and dissimilarity between the texts. For this purpose, we choose surahs containing oaths from Juz' Amma because oath is the common topic between the surahs in this Juz' and work the surahs in a group of three. The analysis is performed in three parts; first is the closeness in terms of distance between the surahs, second is the ordering sequence among the surahs, and third is the closeness against the ordering sequence among the three surahs. The analysis showed that in general, while it is true for smoothness sequence to be based on the traditional closeness, it is not necessarily true for chronological closeness.

Keywords: Stylometry, Chronology, Qur'an, Juz' Amma

1. Introduction

Stylometry is the field which studies the style of text, in order to examine how similar are the texts in terms of styles based on a number of style markers. In literature, stylometry applications are focused on two major applications; the attributional studies and the chronology problems (Holmes, 1998). Stylometry has benefited wide applications in different domains, among which are forensics (Iqbal et al., 2010), Qur'an chronology (Sadeghi, 2011), religious book attribution (Sayoud, 2012), and social network threat detection (Daelemans, 2013). Similar applications are also reflected in the domain of Qur'an.

In solving the chronological problem, the idea is to find similarity or dissimilarity between texts by measuring the distance based on common style markers in the texts such as morphemes (Sadeghi, 2011) or lexical frequency (Nassourou, 2011). This research is based on Sadeghi (2011), who used a set of style markers in the Qur'an called the morphemes in a chronological problem. He calculated the distances between surahs based on the differences of the common morphemes. Instead of using the real chronology order of the surahs in his study, Sadeghi used the traditional order, which is the order used in Qur'an text.

This creates a gap as which order (traditional vs. chronological) is stylistically more similar to one another. Hence, our hypothesis is as follows: The overall stylistic differences share the smoothest sequence and its relation to the surahs orders, whether traditional or chronological. This means that texts or surahs that are close in revelation time (in chronological order), should be close in style. Likewise, surahs that are close in traditional order, should also be close in style.

The remaining of this paper is organized as follows. Section 2 presents a working example for stylistic analysis following Sadeghi (2011). Section 3 presents the analysis based on chronological order on three selected surahs in Juz' Amma. Finally, Section 4 will conclude the paper.

2. Related Work

Before we compare stylistic distance analysis between the two different orderings, we detailed out the steps to perform distance analysis based on the published results by Sadeghi (2011). Note that the working example is solely based on our interpretation to the methods described in the paper.

Sadeghi (2011) measures the distance using the traditional order of surahs based on word morphemes. According to Sadeghi (2011), a morpheme is defined as a word or part of a word. However, in this research, we define morphemes to refer to the set of function words (i.e. prepositions (*w*, *min*), conjunctions letter (*f*), connected noun (*aldhaina*), particles (*ma*, *inna*, *la*), and adverb (*yawm*)). The spatial metaphor distance used in Sadeghi (2011) was based on the four steps. Again, details of the example along these steps are based on our own interpretation on the concept.

2.1. Get the percentage of morphes in the surahs

In his paper, Sadeghi (2011) lists out the percentage but he does not specify the means of of extracting such percentage. We assume the he includes character *w* and *f* even if they are part of a word and not only as a separated preposition or conjunction letter because he defines a morpheme as a word or part of a word.

Table 1 shows the percentage of morpheme in three surahs, which are *Al-Mu'minun* (The Believers), *Al-Ra'd* (Thunder), and *Ya Sin*.

Morpheme	*Al-Ra'd*	*Al-Mu'minun*	*Ya Sin*
w	6.6	5.3	5.7
u	7.2	4.5	4.2
an	1.6	2.4	2.1
fa	1.1	2.7	2.3
llah	1.9	0.7	0.2

Table 1: The percentage of five morphemes in *Al-Mu'minun*, *Al-Ra'd*, and *Ya Sin*

2.2. Get the difference of morpheme(s) among surahs

Because we are working with three surahs, comparison will be carried out in three parts; one pair of surah each time. Table 2 shows the overall measurement for difference in distance between *Al-Mu'minun* and *Al-Ra'd*. Table 3 shows the distance between *Al-Ra'd* and *Ya Sin*. Table 4 shows the distance between *Al-Mu'minun* and *Ya Sin*.

Morpheme	*Al-Ra'd*	*Al-Mu'minun*	Difference
w	6.6	5.3	1.3
u	7.2	4.5	2.7
an	1.6	2.4	0.8
fa	1.1	2.7	1.6
llah	1.9	0.7	1.2
		Distance	7.6

Table 2: Distance between *Al-Mu'minun* and *Al-Ra'd*

Morpheme	*Al-Ra'd*	*Ya Sin*	Difference
w	6.6	5.7	0.9
u	7.2	4.2	3.0
an	1.6	2.1	0.5
fa	1.1	2.3	1.2
llah	1.9	0.2	1.7
		Distance	7.3

Table 3: Distance between *Al-Ra'd* and *Ya Sin*

Morpheme	*Al-Mu'minun*	*Ya Sin*	Difference
w	5.3	5.7	0.4
u	4.5	4.2	0.3
an	2.4	2.1	0.3
fa	2.7	2.3	0.4
llah	0.7	0.2	0.5
		Distance	1.9

Table 4: Distance between *Al-Mu'minun* and *Ya Sin*

2.3. Get the sum of difference in distance

The next step is to calculate the overall measurement of difference in distance, which is essentially the sum of the difference. For example, from Table 2, the sum of difference between *Al-Mu'minun* and *Al-Ra'd* is shown in Equation 1.

$$1.3 + 2.7 + 0.8 + 1.6 + 1.2 = 7.6 \tag{1}$$

Similarly for Table 3 and Table 4, Sadeghi (2011) states the distance between *Al-Ra'd* and *Ya Sin* is 7.3 while the distance between *Al-Mu'minun* and *Ya Sin* is 1.9.

2.4. Analyze the distance to get smoothness sequence

According to Sadeghi (2011), the smaller distance between surahs indicates that they stylistically similar. Based on the analysis from the previous steps, the smoothness sequence that identifies the closeness of the surahs to each other is:

Al-Mu'minun $(1.9) \rightarrow$ *Al-Ra'd* $(7.3) \rightarrow$ *Ya Sin*

Figure 1 illustrates the distance graphically.

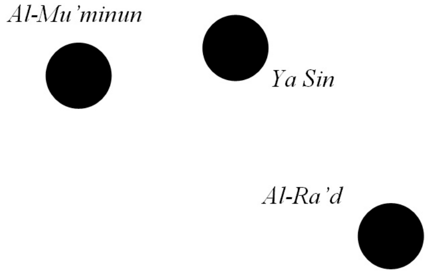


Figure 1: Distance between *Al-Mu'minun* (The Believers), *Al-Ra'd* (Thunder), and *Ya Sin*

3. Stylistic Distance Analysis

There are two types of surahs ordering in the Quranic studies, traditional (based on the chapter ordering in the Qur'an) and chronological (based on revelation date of each surah in the Qur'an). In this paper, we use the chronological order of the surahs to examine the stylistic relationship between surahs whether if the ordering has any effect to the stylistic relationships. We have selected surahs from Juz' Amma (Part 30) as these surahs are very close in terms of subject, direction, and general style (Issa, 2009). To ensure the closeness selection of the surahs, we have only selected the surahs that contain the oath expressions that may start with character *w* or *la uqsim* (Saad et al., 2010).

Following Sadeghi (2011), a distance method can be expressed in Equation 2, which is a general equation to measure the difference between 2 points. Note that the equation is derived from his explanation on how to calculate the distance, but it was not mentioned explicitly in his paper.

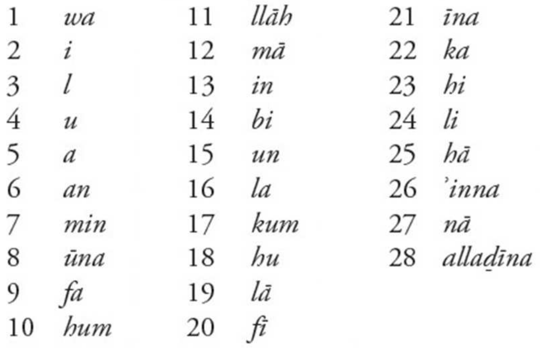
$$\sum_{i=1}^{10}\sum_{j=1}^{8} S_iM_j = S_iM_j - S_{i-1}M_j \tag{2}$$

In this equation, S stands for the 10 surahs in our study and M stands for eight selected style markers (morphemes) in our study.

For the purpose of stylistic distance analysis on Juz' Amma, we have chosen eight morphemes (*w*, *f*, *inna*, *yawm*, *alladina*, *min*, *ma*, *la*) from the list of top 28 frequent morphemes used in Qur'an (Sadeghi, 2011). Figure 2 shows the complete list.

Similar to the example that has been detailed out in the previous section, the overall sum will be calculated by getting the distance between the surahs. To achieve that, we will calculate the difference of each morpheme in the surahs, get the total differences, and then calculate the sum of total differences for every morpheme.

The stylistic distance analysis consists of four steps. In this research, we divide the analysis into three cases; when two surahs are chronologically close while one is not, when all three surahs are chronologically in ordered sequence, and when all three surahs are not entirely in chronologically order sequence but close.



1	*wa*	11	*llāh*	21	*īna*
2	*i*	12	*mā*	22	*ka*
3	*l*	13	*in*	23	*hi*
4	*u*	14	*bi*	24	*li*
5	*a*	15	*un*	25	*hā*
6	*an*	16	*la*	26	*'inna*
7	*min*	17	*kum*	27	*nā*
8	*ūna*	18	*hu*	28	*allaḏīna*
9	*fa*	19	*lā*		
10	*hum*	20	*fī*		

Figure 2: 28 most frequent morphemes in the Qur'an (Sadeghi, 2011)

3.1. Two chronologically close surahs but one is not

The first three surahs under study are *Al-Tariq* (The Night-Comer), *Al-Nazi'at* (The Forceful Chargers), and *Al-Inshiqaq* (Ripped Apart). Table 5 shows the ordering (whether traditional or chronological) as well as the number of words in each surah.

Traditional Ordering	Chronological Ordering	Surah	Number of Words
86	36	*Al-Tariq*	61
79	81	*Al-Nazi'at*	179
84	83	*Al-Inshiqaq*	107

Table 5: Ordering and number of words in *Al-Tariq*, *Al-Nazi'at*, *Al-Inshiqaq*

Next, Table 6 shows the morpheme(s) for all three surahs.

Morpheme	*Al-Tariq*	*Al-Nazi'at*	*Al-Inshiqaq*
w	8	13	11
f	3	16	6
inna	2	6	3
yawm	1	3	0
aldhania	0	0	2
min	2	6	2
ma	4	1	4
la	1	0	3

Table 6: Number of morphemes in *Al-Tariq*, *Al-Nazi'at*, *Al-Inshiqaq*

Based on Table 6, we measured the average weight of each morpheme using the following Equation 3 where w_i is the number of morphemes in the surah and w_n is the total number of words in the surahs.

$$Weight(morpheme) = \frac{\sum w_i}{\sum w_n} \tag{3}$$

Table 7, Table 8, and Table 9 show the average weights for each morpheme across all three surahs. Next, we put all the weights together as shown in Table 10.

Al-Tariq	No. Morpheme (m)	Weight (m/61)
w	6	0.13
f	3	0.05
inna	2	0.03
yawm	1	0.02
alladina	0	0.00
min	2	0.03
ma	4	0.06
la	1	0.02

Table 7: Morpheme weights in *Al-Tariq*

Al-Nazi'at	No. Morpheme (m)	Weight (m/179)
w	13	0.07
f	16	0.09
inna	6	0.03
yawm	3	0.02
alladina	0	0.00
min	6	0.03
ma	1	0.01
la	0	0.00

Table 8: Morpheme weights in *Al-Nazi'at*

Al-Inshiqaq	No. Morpheme (m)	Weight (m/107)
w	11	0.11
f	6	0.06
inna	3	0.03
yawm	0	0.00
alladina	2	0.02
min	2	0.02
ma	4	0.04
la	3	0.03

Table 9: Morpheme weights in *Al-Inshiqaq*

Morpheme	*Al-Tariq*	*Al-Nazi'at*	*Al-Inshiqaq*
w	0.13	0.07	0.11
f	0.05	0.09	0.06
inna	0.05	0.09	0.06
yawm	0.02	0.02	0.00
alladina	0.00	0.00	0.02
min	0.03	0.03	0.02
ma	0.06	0.01	0.04
la	0.02	0.00	0.03

Table 10: Overall weights in *Al-Tariq*, *Al-Nazi'at*, and *Al-Inshiqaq*

The following Equation 4 calculates the difference of each morpheme between a pair of surahs at a time:

$$F(m_i) = m_i S_r - m_i S_k \tag{4}$$

where S represents the surah, m represents the morpheme, and F is the difference in the morpheme. The overall measurement for difference in distance is obtained by summing all the differences as in Equation 5 where D is the difference in distance.

$$D(S_r, S_k) = \sum_{1}^{8} F(m_i) \tag{5}$$

Table 11 in the following page shows the sum of difference for all three surahs *Al-Tariq*, *Al-Nazi'at*, and *Al-Inshiqaq*, respectively. Finally, the smoothness sequence for *Al-Tariq*, *Al-Nazi'at*, and *Al-Inshiqaq* are:

Al-Inshiqaq $(0.11) \rightarrow$ *Al-Tariq* $(0.17) \rightarrow$ *Al-Nazi'at*

Recall that the smaller distance between surahs indicates that they are stylistically similar. The above smoothness sequence shows that *Al-Tariq* and *Al-Inshiqaq* stylistically more similar to one another because they share the same morphological sets. However, in this analysis the smoothness sequence of surahs follows traditional ordering and not the chronological ordering.

3.2. Surahs in ordered sequence

Similarly, we performed the analysis to another set of three surahs that are ordered in sequence. The surahs that we chose are *Al-Shams* (The Sun), *Al-Buruj* (The Towering Constellations), and *Al-Tin* (The Fig). Table 12 shows the ordering as well as the number of words in each surah.

Traditional Ordering	Chronological Ordering	Surah	Number of Words
91	26	*Al-Shams*	54
85	27	*Al-Buruj*	109
95	28	*Al-Tin*	34

Table 12: Ordering and number of words in *Al-Shams*, *Al-Buruj*, *Al-Tin*

The smoothness sequence for *Al-Shams*, *Al-Buruj*, and *Al-Tin* are:

Al-Buruj $(0.14) \rightarrow$ *Al-Tin* $(0.31) \rightarrow$ *Al-Shams*

Figure 3 shows the relative frequency counts of the morphomes in the said surahs. For the analysis of surahs in ordered sequence, we found that the surah *Al-Buruj* is stylistically closer to *Al-Tin* than *Al-Shams*. This is on the contrary with the previous findings whereby similarly in style is detected within traditional order of the surahs. In this case, similarity in style is also detected within the chronological order of the surahs.

3.3. Non-ordered but close surahs

In this section we analyze surahs that are close in distance but not necessarily entirely in chronological order. Table 13 shows the percentage of each morpheme in another three surahs that fit this descriptions, which are *Al-Takwir* (Shrouded in Darkness), *Al-Layl* (The Night), and *Al-Fajr* (Daybreak).

The smoothness sequence for *Al-Takwir*, *Al-Layl*, and *Al-Fajr* are:

Al-Layl $(0.11) \rightarrow$ *Al-Takwir* $(0.17) \rightarrow$ *Al-Fajr*

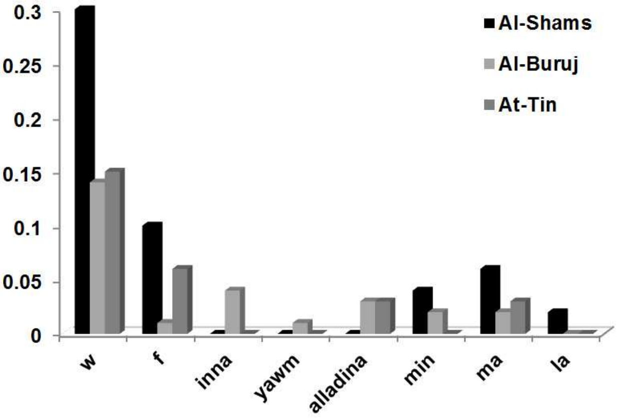


Figure 3: The relative frequency counts of eight morphemes (*wa*, *fa*, *inna*, *yawm*, *aladhia*, *min*, *ma* or *la*) in the surahs of *Al-Shams*, *Al-Buruj*, and *Al-Tin*

Traditional Ordering	Chronological Ordering	Surah	Number of Words
81	7	*Al-Takwir*	104
92	9	*Al-Layl*	71
89	10	*Al-Fajr*	137

Table 13: Ordering and number of words in *Al-Takwir*, *Al-Layl*, and *Al-Fajr*

Figure 4 shows the relative frequency counts of the morphomes in all three surahs; *Al-Takwir*, *Al-Layl*, and *Al-Fajr*.

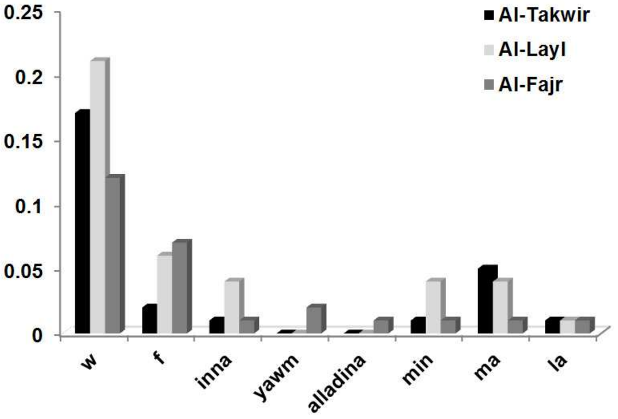


Figure 4: The relative frequency counts of eight morphemes (*wa*, *fa*, *inna*, *yawm*, *aladhia*, *min*, *ma* or *la*) in the surahs of *Al-Takwir*, *Al-Layl*, *Al-Fajr*

Based on the analysis, we found that the surah *Al-Layl* is closer to *Al-Takwir* as compared to *Al-Fajr*. This is inline with the previous findings whereby similarity in style is detected within chronological order of the surahs based on the revelation time, not the traditional order of the surahs.

4. Conclusion

In this paper, we measured stylistic distances of surahs share same topic (i.e. oath) from Juz' Amma in three sce-

Morphemes	Name of Surahs		D	Name of Surahs		D	Name of Surahs		D
	Al-Tariq	*Al-Insyiqaq*		Al-Nazi'at	Al-Insyiqaq		Al-Nazi'at	Al-Tariq	
w	0.13	0.11	0.02	0.07	0.11	0.04	0.07	0.13	0.06
f	0.05	0.06	0.01	0.09	0.06	0.03	0.09	0.05	0.04
inna	0.03	0.03	0	0.03	0.03	0	0.03	0.03	0
yawm	0.02	0	0.02	0.02	0	0.02	0.02	0.02	0
aldhania	0	0.02	0.02	0	0.02	0.02	0	0	0
min	0.03	0.02	0.01	0.03	0.02	0.01	0.03	0.03	0
ma	0.06	0.04	0.02	0.01	0.04	0.03	0.01	0.06	0.05
la	0.02	0.03	0.01	0	0.03	0.03	0	0.02	0.02
	Distance		0.11	Distance		0.18	Distance		0.17

Table 11: Sum of difference for all three surahs *Al-Tariq*, *Al-Nazi'at*, and *Al-Inshiqaq*

narios; surahs that are close to each other (in sequenced chronological order), surahs that are close but has some ordered sequence, as well as two surahs that are chronologically sequenced but one is not. The experiments shown that surahs might be close in style regardless the type of ordering. In summary, what is true for traditional order is not necessary true for chronological order. We can see surahs that are far but close in style.

5. Acknowledgements

This work was supported by Fundamental Research Grants Scheme from the Ministry of Higher Education Malaysia in collaboration with the Centre of Quranic Research at Universiti Malaya, Malaysia.

6. References

W. Daelemans. 2013. Explanation in Computational Stylomery. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2, pages 451–462.

D. I. Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117.

F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi. 2010. Mining Writeprints from Anonymous E-mails for Forensic Investigation. *Digital Investigation*, 7(1):56–64.

A. R. Issa. 2009. *Oath in Juz Amma*. Ph.d. thesis, Darululum College, Cairo University, Egypt.

M. Nassourou. 2011. *A Knowledge-based Hybrid Statistical Classifier for Reconstructing the Chronology of the Quran*. University of Wurzburg, Germany.

S. Saad, N. Salim, Z. Ismail, and H. Zainal. 2010. Towards context-sensitive domain of islamic knowledge ontology extraction. *International Journal for Infonomics*, 3(1).

B. Sadeghi. 2011. The chronology of the qur'an: A stylometric research program. *Arabica*, 58:210–299.

H. Sayoud. 2012. Author Discrimination between the Holy Quran and Prophet's Statements. *Literary and Linguistic Computing*, 27(4):427–444.

Computational ontologies for semantic tagging of the Quran: A survey of past approaches.

Sameer M. Alrehaili, Eric Atwell
School of Computing
University of Leeds, Leeds LS2 9JT, UK
E-mail: salrehaili@gmail.com, e.s.atwell@leeds.ac.uk

Abstract

Recent advances in Text Mining and Natural Language Processing have enabled the development of semantic analysis for religious text, available online for free. The availability of information is a key factor in knowledge acquisition. Sharing information is an important reason for developing an ontology. This paper reports on a survey of recent Qur'an ontology research projects, comparing them in 9 criteria. We conclude that most of the ontologies built for the Qur'an are incomplete and/or focused in a limited specific domain. There is no clear consensus on the semantic annotation format, technology to be used, or how to verify or validate the results.

Keywords: semantics, ontologies, Qur'an

1. Introduction

Recent advances in Text Mining and Natural Language Processing have led to a number of annotations for religious text such as the Qur'an.

Ontology-based models of computational semantics are being widely adopted in various fields such as Knowledge Management, Information Extraction, and the Semantic Web. Religious Studies researchers are also starting to exploit the ontology for improving the capture of knowledge from religious texts such as the Qur'an and Hadith. A definition of ontology in Artificial Intelligence is "the specification of conceptualizations, used to help programs and humans share knowledge.". Ontology development is generally described as an iterative process, and the development process never completes (Ullah Khan et al. 2013). Therefore, many researchers start with a focus on one or two semantic fields of their Qur'an ontology.

There are different annotations and ontologies already available for the Qur'an online and most of them are free. However, they differ in the format that they provide for End-Users, and in the technologies that they use to construct and implement the ontology. This variety of formats used to store the annotated data of the Qur'an leads to a gap between computer scientists, who make tools to provide analysis, and End-Users who are interested in the specific domain. Not all End-Users or linguistics researchers are technically able or willing to make their own converter. Therefore, the need to design a standard format and provide available analyses for the Qur'an in a standard format is becoming essential to facilitate End-User work and make them focus on the analysis instead of doing extra data-reformatting work. Moreover, this would increase the process of development in Qur'an analysis. This paper does not try to solve these challenges, instead of that it tries to survey the Qur'an ontology research projects that have been done recently, comparing them in terms of 9 criteria.

The rest of this paper is organized as follows: section 2 is a brief introduction about the Qur'an. Section 3 identifies the criteria for evaluation of Quran ontologies used in this survey. Section 4 discusses previous work related to the Qur'an and ontology. Finally, conclusion and research future directions are presented. This paper includes a comprehensive comparison table summarising the key features of the different existing ontologies of the Qur'an.

2. The Quran

Muslims believe that the Qur'an is God's word and the most widely read book in the world since its revelation; every Muslim can memorise and recite some parts of the Quran at least 17 times every day when praying. Its recitation and reading have not stopped one day since its revelation. The Qur'an includes a range of knowledge in different subjects such as science, art, stories and history, agriculture and industry, human and social relations, organization of finance, education and health. For some Muslims who do not speak Arabic, and for non-Muslims, the Qur'an is difficult to understand, although it has been translated into over 100 different languages.

3. Comparison criteria

Table 1 summarises the comparison of the Quran ontologies described in the literature. The comparison focuses on the content of ontologies in the work reviewed. The list of criteria used for comparison is described briefly in the following.

1. **Qur'an text:** The ontology relies on one of the following languages:
   - Original Arabic text (A).
   - English translation (B).
   - Malay translation (C).

This criterion has been chosen because we noticed that there is a variation of the language used in ontology-based work. For example (Ullah Khan et al. 2013) and (Saad et al. 2010) ontologies used English translations of the Qur'an, while (Ali & Ahmad 2013) used a Malay translation. This aspect should not be ignored in research on reusing an ontology as it identifies a challenge in merging different translations of Qur'an ontologies.

2. **Coverage area:** Topics and word types that are covered by the ontology. For example, an ontology may covers the topic of animals for only nouns. This aspect compares the ontologies on the topic that they have created for.

3. **Coverage proportion:** This criterion identifies if the ontology covers the entire Qur'an or only some parts.
   - Entire the Qur'an (A).
   - Some parts (B).
   - Specific topic (C).

We found only one work that covers all Qur'an chapters, others focused on one or two topics.

4. **Underlying format:** There are many formats such as plain text, XML files, and RDF or OWL.

Format is also an important factor in ontology reuse due to requirements for extra work in extraction of ontology elements from the existing ontologies.

5. **Underlying technology used:** tools used for building and representing the ontology.

6. **Availability:** this criterion identifies if the ontology is free access or not. This is important in reusing an ontology too because we have noticed that there are some resources which are not available for download and reuse.
   - Yes (A)
   - No (B)

7. **Concepts number:** The number of abstract and concrete concepts in the ontology.

8. **Relations type:** The ontology may be have one of the following relations between the concepts:
   - Meronymy (Part-of) (A)
   - Synonyms (similar) (B)
   - Antonymy (opposite) (C)
   - Hyponymy (subordinate) (D)

9. **Verification method used:** The evaluation method used to verify the ontology. Two types of methods have been used to verify ontology-based work on the Qur'an:
   - Domain experts
   - Scholarly sources (Ibn Kathir)

This gives us information about quality of the work that has been conducted in order to evaluate the ontology.

4. Ontology research on the Qur'an

Several initial studies have been undertaken on the topic of Qur'an ontology. Most of these studies have been developed in order to improve the efficiency of information retrieval for the Qur'an. These have facilitated the process of accessing Qur'an knowledge. However, they vary from each other in different aspects such as coverage of the Qur'an, discourse level, language of the text used; original Arabic text or other translation, domain focused on, number of concepts and types covered, concept extraction method, relation types they provide, development process they followed during construction, technology used in ontology construction, availability, and verification method.

(Saad et al. 2009) proposed a simple methodology for automatic extraction of a concept based on the Qur'an in order to build an ontology. This paper used a method based on extracting the verses which contain a word of prayer in it as well as the previous and next verse. This method relies on a format of one English translation of the Qur'an that included some aspects such as Uppercase Letter. An uppercase letter is used to identify the concepts such as the Book. Another feature called compound noun is used to identify the relationship of hyponym or "Part-OF" between the concepts. A copula is used to identify the syntactic relationship between subject and adjective or noun. The ontology is based on the information obtained from domain experts. The development process is adopted from (Saad & Salim 2008). However, the authors have focused on the subject of Prayer or "Salat" particularly in daily prayer, thus this ontology does not cover all subjects in the Qur'an. In addition, there is no mention about underlying format or ontology technologies used in this paper.

Saad et al. continued their work to develop a framework for automated generation of Islamic knowledge concrete concepts that exist in the holy Qur'an as presented in (Saad et al. 2010). The framework takes into account some situations form the sciences of the Qur'an, such as the cause of revelation (Asbab Al Nuzul), and verses overridden by related verses that were revealed later (Nasikh Mansukh). The methodology of ontology development was also adopted from (Saad & Salim 2008), and the method to obtain the concepts is applying a grammar and extraction rules to the English translation of the Qur'an. The 374 extracted instances only cover verses that have the keyword salah or pray and this does not cover the entire Qur'an. These instances were mapped to six abstract concepts. This paper differs from the previous in synonym relations.

(Saad et al. 2011) proposed methods for designing an ontology based on translated texts of the Qur'an. Information used in developing the ontology was collected by the domain experts. Their ontology also only covers the subject of "Salat" (pray).

Another ongoing research project on a prototype of a framework called SemQ is presented in (Ai-khalifa & Ai-yahya 2009). SemQ identifies opposition relationships between Quranic concepts. The idea is SemQ receives a verse as input and produces a list of words that are opposed to each other with the degree of the opposition. The coverage is in the domain of “Women” in the Qur’an. Ontology development makes use of the Buckwalter morphology POS annotation and focuses on nouns and verbs that are related to the semantic field of Time. This paper used OWL and UPON technologies in order to represent the concepts and relations. The ontology consists of seven abstract concepts and eleven concrete concepts. This ontology is sharable and can be downloaded. This study was limited to word level which includes only nouns and verbs of the “Women” domain. However, there are no evaluable results provided by the authors or any validation attempts for their proposed framework.

In (Ali & Ahmad 2013) , a theme-based approach is proposed to represent and classify the knowledge of the Qur’an using an ontology. Their ontology was developed according to themes described in Syammil Al-Quran Miracle the Reference, and using protégé-OWL and Malay language as medium of concepts, and was validated by the domain experts. It only covers two themes: “Iman” which means faith and “Akhlaq” which means deed. This was an Ontology-based approach to represent and classify Qur’anic concepts according to specific semantic fields. The structure of the ontology was verified by Qur’an domain experts. The ontology was developed using Protégé-OWL and using Malay Language as the medium language. The authors proposed a representation approach whcih differs from traditional representation which consist of Juz, Chapter and Verse. There is no explanation of what language was used for this ontology and what source the concepts were based on. They implemented the ontology using protégé. There are no details of results or validation of the ontology, although the paper states that the process of creating the ontology was reviewed by seven Qur’an domain experts.

(Ullah Khan et al. 2013) developed a simple ontology for the Qur’an that includes the animals that are mentioned in the Qur’an in order to provide Qur’anic semantic search. The ontology was built using protégé editor, and SPARQL query language was used to retrieve the answers to a query. The English translation of the Qur’an by Pickthall is used in this ontology. The ontology provides 167 direct or indirect references to animals in the Qur’an obtained based on information mentioned in a book entitled “Hewanat-E-Qurani”. The relationship type is a taxonomy relation. The paper concludes that the existing Arabic WordNet does not help for retrieving this type of document information.

(Yauri et al. 2012) proposed a model for defining the important Qur’anic concepts by knowledge representation and presented the relationships between them using Description Logic (Predicate logic). They reused the Quranic Arabic Corpus ontology by (Dukes 2013). This ongoing research attempts to reuse and improve an existing ontology developed in Leeds by adding more relations. Protégé is used in ontology construction. A top-down ontology development process was followed. It has 15 abstract concepts.

(Yauri et al. 2013) has proposed ontology-based semantic search for the Qur’an using protégée editor and Manchester OWL query language. The ontology was built by reusing the existing Quranic Arabic Corpus ontology developed by (Dukes 2013), and adding more than 650 relationships depending on the Qur’an, Hadith, and Islamic websites. This ontology was constructed manually.
(Yahya et al. 2013) proposed a semantic search for the Qur’an based on Cross Language Information Retrieval (CLIR). They created a bilingual ontology for the Quran composed of concepts based on existing Quranic Arabic Corpus ontology by (Dukes 2013), and found 5695 documents belonging to a main concept, where 541 documents are not assigned to any concepts in an English translation. In Malay, there are 5999 documents assigned to main concepts, where 237 documents do not belong to any concept.

In (Yunus et al. 2010), the authors did experiments on retrieving verses of the Quran using a semantic query approach exploiting Cross Language Information Retrieval (CLIR).

(Abbas 2009) developed a tool for searching for the Qur’anic concrete and abstract concepts. She exploited an existing Qur’an topics index from a scholarly source: Tafsir of Ibn Kathir. This onotology covered the whole Qur’an.

(Dukes 2013) in his PhD thesis defines 300 concepts in the Qur’an, and extracts the interrelationships using Predicate logic. The number of relations is 350. The type of relation between concepts is Part-of or IS-A. The ontology is also based on the Tafsir by Ibn Kathir.

(Muhammad 2012) in his thesis, developed an ontology covering the whole Qur’an in terms of pronoun tagging. Each pronoun is linked to its antecedent or previous subject.

In (Sharaf & Atwell 2012), the authors have created a dataset called QurSim which consists of 7600 pairs of related verses for evaluating the relatedness of short texts.

An automatic knowledge extraction method based on rules and natural language patterns is described in (Saad et al. 2013). Their methods rely on the English translation of the Qur’an and have identified a new pattern language named Qpattern which is suitable for extraction of taxonomy part-of relations. This research also identified that it is difficult to extract information from text that includes co-reference like the Qur’an.

The aim of this report was to look at the range of existing studies on Quran ontology available currently and identify the limitations of these studies as well as potential future work. Some semantic annotations have been done for the entire Qur'an, but for a specific type of word and domain, such as (Al-khalifa & Al-yahya 2009), an ontology for verbs in the domain of women, or the ontology of (Al-yahya et al. 2010) for nouns in the domain of time. There is one non-domain-specific ontology for the entire Qur'an but it is only for pronouns (Muhammad 2012). Most ontologies have relations using Part-Of or synonyms, but one work includes opposition relations, (Ai-khalifa & Ai-yahya 2009). Most ontologies built for the Qur'an are incomplete and focused in a specific domain. There is no clear consensus on the semantic annotation format, technology to be used, or how to verify or validate the results.

5. Acknowledgements

Sameer M. Alrehaili is supported by a PhD scholarship from the Ministry of Higher Education, Saudi Arabia. We thank the 2 reviewers for their comments on this paper.

6. References

Abbas, N., 2009. Quran'search for a Concept'Tool and Website. *MRes Thesis, School of Computing, Univeristy of Leeds.*

Al-khalifa, H. & Al-yahya, M.M., 2009. SemQ : A Proposed Framework for Representing Semantic Opposition in the Holy Quran using. *Current Trends in Information Technology (CTIT), 2009 International Conference on the*, pp.0–3.

Ali, B. & Ahmad, M., 2013. AL-QURAN THEMES CLASSIFICATION USING ONTOLOGY. *icoci.cms.net.my*, (074), pp.383–389. Available at: http://www.icoci.cms.net.my/proceedings/2013/PDF/PID74.pdf [Accessed November 26, 2013].

Dukes, K., 2013. *Statistical Parsing by Machine Learning from a Classical Arabic Treebank*, PhD Thesis, School of Co mputing, University of Leeds

Muhammad, A.B., 2012. *Annotation of Conceptual Co-reference and Text Mining the Qur'an*, PhD Thesis, School of Computing, University of Leeds.

Saad, S. et al., 2010. A framework for Islamic knowledge via ontology representation. *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*, pp.310–314. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5466897.

Saad, S. & Salim, N., 2008. Methodology of Ontology Extraction for Islamic Knowledge Text. In *Postgraduate Annual Research Seminar, UTM.*

Saad, S., Salim, N. & Zainal, H., 2009. Pattern extraction for Islamic concept. *Electrical Engineering and …*, (August), pp.333–337. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5254719 [Accessed November 26, 2013].

Saad, S., Salim, N. & Zainal, H., 2013. Rules and Natural Language Pattern in Extracting Quranic Knowledge. In *In the proceedings of the Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences (NOORIC 2013).* Madinah, Sadudi Arabia: IT Research Center for the Holy Quran and Its Sciences (NOOR).

Saad, S., Salim, N. & Zainuddin, S., 2011. An early stage of knowledge acquisition based on Quranic text. *Semantic Technology and …*, (June), pp.130–136. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5995777 [Accessed November 26, 2013].

Sharaf, A.-B. & Atwell, E., 2012. QurSim: A corpus for evaluation of relatedness in short texts. In N. C. (Conference Chair) et al., eds. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey: European Language Resources Association (ELRA).

Ullah Khan, H. et al., 2013. Ontology Based Semantic Search in Holy Quran. *International Journal of Future Computer and Communication*, 2(6), pp.570–575. Available at: http://www.ijfcc.org/index.php?m=content&c=index&a=show&catid=43&id=493 [Accessed December 16, 2013].

Yahya, Z. et al., 2013. Query Translation Using Concepts Similarity Based on Quran Ontology for Cross-Language Information Retrieval. *Journal of Computer Science*, 9(7), pp.889–897. Available at: http://thescipub.com/abstract/10.3844/jcssp.2013.889.897 [Accessed November 7, 2013].

Yauri, A.R. et al., 2013. Ontology Semantic Approach to Extraction of knowledge from Holy Quran. *Computer Science and Information Technology (CSIT), 2013 5th International Conference*, pp.19–23. Available at: http://0-ieeexplore.ieee.org.wam.leeds.ac.uk/stamp/stamp.jsp?tp=&arnumber=6588752&isnumber=6588741.

Yauri, A.R. et al., 2012. Quranic-based concepts: Verse relations extraction using Manchester OWL syntax. *2012 International Conference on Information Retrieval & Knowledge Management*, pp.317–321. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6204998.

Yunus, M. a, Zainuddin, R. & Abdullah, N., 2010. Semantic query for Quran documents results. In *2010 IEEE Conference on Open Systems (ICOS 2010).* Ieee, pp. 5–7. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5719959.

Reference	1	2	3	4	5	6	7	8	9
(Saad et al. 2009)	B	“Pray”	C	N/A	N/A	N/A	N/A	A, Part-Of	Domain experts
(Saad et al. 2010)	B	“Pray”	C	N/A	N/A	N/A	374 instances and 6 abstract	B, synonyms	Domain experts
(Ai-khalifa & Ai-yahya 2009)	A	“Women”, Nouns and Verb	C	OWL	UPON	A	11 concrete and 7 abstract	C, opposition	N/A
(Al-yahya et al. 2010)	A	“Time noun”	C	OWL	UPON	A	11 concrete and 7 abstract	D, hyponymy	N/A
(Ali & Ahmad 2013)	C	“faith and deed”	C	OWL	protégé	N/A	N/A	N/A	Domain experts
(Ullah Khan et al. 2013)	B	“animals”	C		Protégé, SPARQL	N/A	N/A	A, Part-od	N/A
(Yauri et al. 2012)	N/A	”salat, zakat, sin, reward”	C	OWL	Protégé	N/A	15 abstract	N/A	N/A
(Yauri et al. 2013)	N/A	N/A	N/A	Manchester OWL	N/A	N/A	N/A	N/A	Manually constructed
(Yahya et al. 2013)	B, C	N/A	C	N/A	N/A	N/A	N/A	N/A	N/A
(Yunus et al. 2010)	A, B, C	N/A	C	N/A	N/A	N/A	N/A	N/A	N/A
(Dukes 2013)	A, B	N/A	N/A	Text files	N/A	B	300	Part-of	Ibn Kathir
(Muhammad 2012)	A	pronouns	A	XML	N/A	A	N/A	N/A	Ibn Kathir

Table 1: summary of ontology features in papers reviewed

Combining Critical Discourse Analysis and NLP tools in investigations of religious prose

Bruno Bisceglia, Rita Calabrese, Ljubica Leone*
University of Salerno
Department of Industrial Engineering, Department of Humanities
E-mail: bbisceglia@unisa.it, rcalabrese@unisa.it, lleone@unisa.it

Abstract

The present paper aims to investigate the discourse strategies adopted in selected samples of religious prose with the aim to: 1. Identify the main text functions which may result in differing subsections of the texts in terms of exhortation, exposition, narration and argumentation; 2. Compare the main principles and methodological procedures of Critical Discourse Analysis (CDA) with the corpus data. CDA explores the relationship between the exercise of power and discourse structuring by looking at the structures of power, authority and ideology which underlie the structures of speech and writing. To verify the above assumptions, a corpus of documents released by the Second Vatican Council (1962-65) was collected and automatically parsed by using the language analysis tools available at the Visual Interactive Syntax Learning (VISL) website. Along with the primary corpus, a smaller control corpus of documents issued by the First Vatican Council in 1869-70 was created and annotated to provide a comparable basis to the analysis of the data. Following the automatic procedure of detection and extraction of information from a parsed corpus, we have matched corpus-based evidence and linguistic diagnostics (e.g. modality) generally used in CDA to determine the level of recursion and innovation, authority and liberality characterizing our data.

Keywords: CDA, modality, parsed corpus

1. Motivation

The present paper aims to investigate the discourse strategies adopted in the documents released by the Vatican Council II in order to: 1. Identify the main text functions which may result in differing subsections of the texts in terms of exhortation, exposition, narration and argumentation; 2. Compare the main principles and methodological procedures of Critical Discourse Analysis with the corpus data. The analysis of the corpus could contribute to a better understanding of the documents either in terms of a theological discussion which involves expert members of the Christian community addressing other specialists/theologians or a theological discussion actually addressing ordinary lay people. In one case, the religious discourse should result in the use of redundant and recurrent features typical of the traditional religious prose; in the other, it should contain divergent and innovating elements regenerating not only the prose but also the inner spirit of the ecumenical message.

We assume that an analysis of religious texts based on the above assumptions could provide important insights into the interpretation of the documents under study as well as basic methodological directions for future investigations into specialized texts in general and religious texts in particular.

Critical Discourse Analysis (CDA) is a recently developed approach to text analysis (Fairclough 2010; Rogers 2004) which examines the role of language in classifying phenomena and experiences through which individuals interpret reality assuming that the use of language is part of a wider ideological process of meaning construction. In particular, it explores the relationship between the exercise of power and discourse structuring by looking at the structures of power, authority and ideology which underlie the structures of speech and writing. The term 'critical' used in the approach refers to the active role assigned to individuals who are encouraged to question assumptions rather than taking them for granted (Clark 2007, p. 138). Under this view, religious language is often subjected to controversial debates since it is considered as both a powerful source of language rule-breaking and language bending in its effort to firmly express the sense of something that exists beyond common language and real world: "The devising of new ways of talking about God is always a controversial activity, given the conservative forces within religious expression" (Crystal 2010, p.403).

To verify the above assumptions, a corpus of documents produced during the Second Vatican Council (1962-65) was collected and automatically parsed by using the language analysis tools provided by the Visual Interactive Syntax Learning (VISL) website[1] which can provide both syntactic and semantic information on a given constituent structure. Along with the primary corpus, a smaller control corpus of documents issued by the First Vatican Council in 1869-70 was created and annotated to provide a comparable basis to the analysis of the data. Following the automatic procedure of detection and extraction of information from a parsed corpus, we have matched corpus-based evidence and linguistic diagnostics (e.g. modality) generally used in CDA to determine the level of recursion and innovation, authority and liberality underlying/characterizing our data.

The paper is organized as follows. Section 2 provides an historical outline of the council along with essential information about the content of its documents. Section 3

[1] http://beta.visl.sdu.dk/visl/en/parsing/automatic/

introduces the theoretical framework in which the study can be set. Section 4 shows the results of a study of those language features which are generally considered as characterizing ideological and authoritative discourse, namely modality and transitivity, with specific reference to our corpus. Finally, the implications of the present findings for future research on specialized language will be discussed in section 5[2].

2. The Second Vatican Council

The Second Vatican Council (1962-65) was the 21st ecumenical council of the Roman Catholic Church announced by Pope John XXIII in 1959. Some preparatory commissions were appointed by the Pope with the aim to prepare the agenda of the council and provide drafts (schemata) of decrees on various topics. On the opening of the council in 1962, the council fathers coming from various parts of the world were recommended to accomplish the pastoral duties of the church.

After the Pope's death, the work of the council was carried on by his successor, Paul VI, and later completed in 1965. At the end of the sessions, sixteen documents were promulgated and divided into three different categories: constitution, declarations and decrees. The original documents were written in Latin and then translated into thirteen languages including non-European languages such as Swahili and Hebrew.

2.1 The Second Vatican Council Documents

2.1.1 Dei Verbum. Dogmatic Constitution on Divine Revelation

The "Dogmatic Constitution on Divine Revelation" attempts to relate the role of Scripture and the tradition of the postbiblical teaching in the Christian Church to their common origin in the Word of God. The document claims the value of Scripture for the salvation of mankind and maintains an open attitude towards the scholarly study of the Bible.

2.1.2 Lumen Gentium. Dogmatic Constitution on the Church

The "Dogmatic Constitution on the Church" represents the attempt made by the council fathers to use biblical terms rather than juridical categories to describe the organisation of the Church. The discussion about the hierarchical structure of the Church counterbalances to some extent the monarchical role of the papacy as it was intended by the first Vatican Council's teaching. In this way, bishops are given a new and more significant role is in the Christian community.

2.1.3 Gaudium et Spes. Pastoral Constitution on the Church in the World of Today

The "Pastoral Constitution on the Church in Today's World" acknowledges the profound changes mankind is experiencing and attempts to relate both the ideas of Church and Revelation to the needs and values of contemporary culture.

2.1.4 Sacrosanctum Concilium. Constitution on the Sacred Liturgy

The "Constitution on the Sacred Liturgy" establishes the principle of a larger participation of laypeople to the celebration of mass and authorizes significant changes in the traditional texts, forms, and language used during the celebration of mass and the administration of sacraments.

2.1.5 Decrees deal with practical questions such as:

- *Christus Dominus* on the pastoral duties of bishops.
- *Unitatis Redintegratio* on ecumenism.
- *Orientalium Ecclesiarum* on the Eastern-rite churches.
- *Presbyterorum Ordinis* on the ministry and life of priests.
- *Optatam Totius* on education of priests, religious life, the missionary activity of the Church.
- *Apostolicam Actuositatem* on the apostolate of laity.
- *Inter Mirifica* on the role of media in social communication.
- *Perfectae Caritatis* on renewal and readaptation of religious life.
- *Ad Gentes* on church's mission in the world.

2.1.6 Declarations are concerned with general issues such as:

- *Dignitatis Humanae* on religious freedom.
- *Nostra Aetate* on the Church's attitude towards non-Christian religions.
- *Gravissimum Educationis* on Christian education.

The documents of the Council reflect the changes affecting various areas of the Church's life, ranging from biblical, ecumenical and liturgical issues to the long-questioned debate concerning the lay apostolate. They represent therefore an ongoing process started some decades earlier by Pope John. Since then, from the early 1970s on, the contents of the documents and the council's deliberations have largely influenced any field of the Church's life and have initiated important changes still visible at the present time in the pastoral path undertaken by Pope Francis[3].

3.1 Theoretical background

CDA is an interdisciplinary approach to the analysis of social discourse rooted in a specific model of linguistic theory known as systemic functional analysis which is concerned with the way in which linguistic structures are

[2] *Bruno Bisceglia, sj, Department of Industrial Engineering, is author of the abstract and of section 2; Rita Calabrese, Department of Humanities, is author of sections 1 and 3; Ljubica Leone, Department of Humanities, is author of sections 4 and 5.

[3] See for example Pope Francis' message for the 48th World Communications Day available at http://www.vatican.va/holy_father/francesco/messages/communications/documents/papa-francesco_20140124_messaggio-comunicazioni-sociali_en.html

related to communicative functions and social values. In particular, the ideational function pertains the ways language is used to represent the experiential world, whereas the interpersonal function is about how speakers orientate and shape their utterances as discourse (Simpson 2004, p.123). The former is primarily captured by the grammatical system of transitivity which encodes actions and events as well as thoughts and perceptions into the grammar of the clause (ib. p.22), the latter is expressed principally by the system of modality which allows to attach expressions of belief, attitude and obligation to what is said and written. In traditional grammars, the term transitivity is used to identify verbs which take direct objects, whereas in the extended semantic sense provided by a functional model of transitivity, it includes the types of process (classified as material, mental, behavioural, verbal, relational and existential processes) represented in language and the types of participant (agent, goal, patient/receiver) associated with those processes. Modality[4], in a wider sense, is to be intended as the grammar of explicit comment that includes the varying degrees of certainty and of commitment or obligation (Simpson 2004, p.123). In particular, Fowler (1996) identified four main aspects of modality: truth, obligation, desirability and permission. Truth modality is expressed on a cline ranging from absolute confidence (*will*) to uncertainty (*could/might*), in other words from positive (also known as deontic modality) to negative shading (also termed as epistemic modality). Therefore, each modal can convey two different types of meaning which Biber et al. (1999, p.485) labelled as 'intrinsic' (or deontic) and 'extrinsic' (or epistemic). Two typical structures correlate modals with intrinsic meanings: 1. The subject of the verb phrase usually refers to a human being as agent, 2. The main verb is usually a dynamic verb describing an activity that can be controlled by the subject. In contrast, extrinsic meanings are related to modal verbs occurring with non-human subjects and stative verbs. This model of language shows how "points of view and beliefs are not only socially constructed but also grammatically encoded in language in ways that often reveal unequal relationships based on authority and power. [Therefore] language as primary medium of social control and power is fundamentally ideological" (Clark 2007, p.153). For this reason, we think that an analysis of religious texts based on the above assumptions could provide important insights into the interpretation of the documents under study as well as basic methodological directions for future investigations into specialized texts in general and religious texts in particular.

3.2. The present study

The study builds on the above assumptions and aims to investigate discourse strategies adopted in the Vatican documents in order to: 1. Identify the main text functions which may result in differing subsections of the texts in terms of exhortation, exposition, narration and argumentation; 2. Compare the main principles and methodological procedures of CDA with the corpus data. The analysis of the corpus could contribute to a better understanding of the documents either in terms of a theological discussion involving expert members of the Christian community addressing other specialists/theologians or a theological discussion actually addressing ordinary lay people. In one case, the religious discourse should result in the use of redundant and recurrent features typical of traditional religious prose; in the other, it should contain divergent and innovating elements regenerating not only the prose but also the inner spirit of the ecumenical message. Along with the primary corpus, a smaller control corpus of documents issued by the first Vatican Council in 1869-70 was created and annotated to provide a comparable basis to the analysis of the data. The collection we used in the study is entitled *Acta et decreta sacrosancti oecumenici concilii Vaticani in quatuor prionbus sessionibus* and was issued in 1872[5].

Given the overall differences concerning size and time periods covered by the two corpora, the documents of the first Council were compared to the subcorpus of the declarations contained in the second Council corpus due to similarities in style, topic and purpose characterizing both corpora.

3.2.1 Method

In order to address the above issues, we have matched corpus-based evidence and linguistic diagnostics, namely verb phrases with modal auxiliaries as their heads, in order to explore linguistic features that are functionally related and relevant to religious prose and establish the extent to which the frequency of such features across the corpus may contribute to the identification of a specialized and ideologically authoritative prose.

3.2.2 Materials

The data collected in both (sub-) corpora include different complete texts in the form of decrees, declarations and constitution that are however similar in terms of medium (writing), genre (argumentative) and field (public), even though they display some differences in size (each containing respectively 12.494, 81.195, 76.932 tokens) and consequently in the number of occurrences of the linguistic features under study.

3.2.3 Procedure

Once collected, the data have been automatically parsed by using the language analysis tools provided by the VISL interface. The parsers provided by VISL are based

[4] Modality is mostly expressed by modal auxiliary verbs, but lexical verbs, adverbs, adjectives and even intonation may assume a modality connotation. In the present paper, we have restricted our analysis only to modal verbs.

[5] This version is available at http://www.ewtn.com/library/councils/v1.htm#7 and corresponds to the translation of the decrees published by Norman Tanner in 1990.

on *Constraint Grammar* (CG), a methodological paradigm for Natural Language Processing (NLP) (Karlsson et al.1995; Bick 2000) which includes context dependent rules that assign grammatical tags to tokens in a given corpus. Its tags address lemmatisation (lexeme or base form), inflexion, derivation, syntactic function**,** dependency, valency, case roles, semantic type. The system also marks the *dependency relation structures* between parts of speech (POS) with the symbol @ placed before (>) or after (<) the head. Upper case tags describe word classes as well as morphological inflections (e.g. MV= main verb, PRP= preposition, N = noun, GN= genitive). In example (1) the infinitive verb (V INF) is annotated as right argument (@ICL-AUX<) of the present tense (PR) auxiliary/modal <aux> (V) occurring in the main clause @FS-STA.

(1) particular [particular] ADJ POS @>N autochthonous [autochthonous] ADJ POS @>N churches [church] N P NOM @SUBJ> should [shall] <aux> V IMPF @FS-STA be [be] <mv> V INF @ICL-AUX sufficiently [sufficiently] ADV @>A established [established] ADJ POS @<SC #18->16 and [and] KC @CO should [shall] <aux> V IMPF @FS-STA grow [grow] <mv> V INF @ICL-AUX< up [up] ADV @MV< all over [all=over] PRP @<ADVL the [the] ART S/P @>N world [world] N S NOM @P<

Once annotated, tags/instances for each feature could automatically be extracted from the corpus with the application of the *ConcApp* concordancer and then manually mapped onto the corresponding structural patterns selected for the study. In order to accurately classify and estimate all @AUX occurrences in the annotated corpus, we established a specific syntactic setting in our queries and we could extract all examples of @AUX and their corresponding right/left collocates. In particular, we left out from the analysis all semi-modals like *ought to*, *need to* and periphrastic structures like *have to*, *have got to* and the like and focused on the modal patterns listed below:

1. modal + main verb/ VP
2. modal + perfect HAVE + main verb/ VP
3. modal + passive BE + main verb/ VP
4. modal + perfect HAVE + passive BE + main verb/ VP

4. Results and Discussion

The texts included in the corpus show a different number of tokens, therefore it was necessary to calculate the normalized frequency (*Nf*) in order to make them reciprocally comparable. Table 1 below shows *can, must* and *should* as the most frequent modals across the corpus, whereas *shall* and *could* share a quantitatively lower position, probably because they can be replaced by *will* and *can* respectively with a slight or no change in meaning. Likewise, *might* shows a lower frequency when compared to other modals in both corpora and this can be related to its limited use in formal contexts, where it is often replaced by *may*. As for the distinction between *shall* and *will* it must be taken into account that, even though they both express volition or prediction (Quirk et al.1985), they cannot be used in an interchangeable way.

In each subcorpus, however, the most frequent verb is *should*, and this is probably related to the fact that "its force makes it less face-threatening in deontic use" (Aarts et al. 2013, p.79) in comparison with *must* that shares with *should* the same semantic as well as functional field of obligation / necessity. As regards the modal verb *would*, it is worth noting that its average low frequency is quite the same in all subcorpora, probably because it expresses tentativeness rather than strong obligation, a meaning that is unlikely to occur in "normative" texts like the documents of the Council. As a matter of fact, they are to be interpreted to some extent as general rules for both the Church and the whole Christian community.

The frequency list of core modals and their distribution across the documents are shown in the table below along with their raw/normalized frequency (Rf/Nf respectively).

CORE MODALS	Declarations		Constitution		Decrees	
	Rf	Nf	Rf	Nf	Rf	Nf
can	20	16.0	207	25.5	156	20.3
could	-	-	8	0.9	12	1.5
may	19	15.2	208	25.6	210	27.3
might	1	0.8	48	5.9	50	60.5
must	21	16.8	505	62.2	155	20.1
shall	-	-	45	5.5	10	1.3
should	31	24.8	234	28.8	647	84.1
will	14	11.2	172	21.2	185	24.0
would	3	2.4	34	4.2	21	2.7

Table 1. The raw frequency and the normalized frequency (per 10,000 words) of core modals in the Council corpus

As for the semantic aspects in the use of modals, i.e. their extrinsic (epistemic) and intrinsic (deontic) meaning, all @AUX were searched for in the corpus with reference to human vs. non human subjects (NOM) and the main verbs (MV) which follow them. Each subcorpus contains verbs belonging to both categories, except for *shall* and *could* that are not attested in the Declarations, but still present in the first Council documents (*shall*=11 and *could*=2) where both modals convey a deontic rather than epistemic meaning. Looking at the normalized frequency (*Nf*), it seems that there is a general trend affecting each subcorpus with a special predominance of intrinsic meaning with a high number of instances of *may* and *should*, whereas *shall* and *could* show weak use, and the same can be said for *would*. The relatively low frequency of modals conveying an extrinsic meaning is a surprising element in contrast with religious texts that traditionally express rules, advice and prescriptive norms. At the same time, such findings reveal the underlying principles of the council clearly supporting the strength of human willingness and man's freedom of choice in everyday life. What emerges at a first glance is that the subcorpora share some similarities in the distribution of both intrinsic and extrinsic uses of modals, with the exception of *should* that shows its highest frequency in the Decrees and the Declarations. In the Constitution, instead, the frequency of *should* has a downturn to 8,2 and 9,7 (*Nf*), with reference to intrinsic and extrinsic meanings respectively. The modals that show lower frequency in the data are *shall, could* and also *would* and *may* in the Declarations, that can be explained by the overall low frequency of these auxiliaries across the data. The total number of intrinsic and extrinsic modals per subcorpus is shown in Tables 2 and 3.

	Declarations		Constitution		Decrees	
	intr	extr	intr	extr	intr	extr
can	6,4	3,2	10,0	7,2	6,8	7,0
could	-	-	0,5	0,5	0,5	0,5
may	10,4	-	8,0	7,5	11,4	6,2
might	-	0,8	4,7	1,1	4,3	1,3
must	6,4	4,0	8,0	7,0	6,5	5,0
shall	-	-	3,0	2,4	0,6	0,1
should	17,6	20,0	8,2	9,7	36,1	20,4
will	4,0	4,8	7,0	6,0	9,6	10,0
would	-	1,6	3,3	0,6	0,9	0,9

Table 2. Intrinsic vs. extrinsic modality per subcorpus (*Nf* per 10,000 words)

Modals	*n*= Deontic	*n*= Epistemic	*TOT*
can	5	20	25
could	2	-	2
may	6	10	16
might	6	-	6
must	2	-	2
shall	7	4	11
should	4	14	18
will	8	-	8
would	2	-	2

Table 3. Intrinsic vs. extrinsic modality in the first Vatican Council corpus (*Rf*)

A comparative analysis of the data in both subcorpora shows a predominant assertive meaning co-occurring with stative verbs such as *be* and a high number of passive forms (*n*=202) which seem to be the most striking features of the first Council documents. It is also worth noting the lack of *might, must, would, will* and *could* with extrinsic / epistemic meaning clearly explained and counterbalanced by the higher number (*n*=49) of modals expressing the logical status of events (permission, necessity, prediction) reported in the texts.

As regards the intrinsic sense of modals, something else must be added and in particular some considerations are to be made with reference to the lexical associations with specific subjects and lexical verbs that usually follow them. According to Biber et al. (1999, p.485) "there are two typical structural correlates" of modals with intrinsic meaning, that is the presence of humans responsible for the actions and the verbs that follow them which usually convey a dynamic meaning. As for the former point, it is worth underlining that the most frequent linguistic items functionally behaving as subjects, are pronouns, mainly *he*, *they*, but also nouns like *man*, *men*, *All* (with reference to mankind), and *people* in general. Another important aspect, even though not very frequent in the corpus, is also represented by the use of terms belonging to the domain of religion and playing the functional role of subjects within the clause: for example, *incarnate word*, *God*, *creature*, *Son of God*, *pastor*, *Christians*. Some examples are shown in (2) and (3) below.

(2) Christians will provide, on behalf of family, those necessity. (Constitution)
(3)God will render their deserts to all those…(Declarations)

It is also significant the use of collective nouns such as *Church*, in the sense of community of people able to think and act*, nations*, *council*, *government*, as shown in (4) and (5).

(4) The council could now direct the attention of all…(Constitution)
(5) Church should enjoy that full measure of freedom…(Declarations)

As for the main verbs that follow the modals, a general preference for non stative verbs is attested across the two corpora. This finding can be explained as a general characteristic of the intrinsic sense of modals followed by dynamic verbs that semantically express activity or events such as: *direct, cooperate, entrust, form, give, shine*. This feature draws on a general attitude towards the concept of Christian community characterized by a marked even though gradual shift from idealized principles of Christianity to a renovated process of comprehension, openness and dialogue.

5. Conclusion

The "authors" of the Council's documents aimed at expressing specific directions by placing individuals in the middle of the universe, in a philanthropic view that sets humans in higher ranks as morally responsible agents without distinction of religion (it is important the reference to non Christians in the texts) and social status. The Council aimed at preserving the tradition as well, but rather than simply reiterating old traditions, it first tried to actualize them and, in the light of the changing events of that time, to give them a more pragmatic interpretation. The Council did not intend to develop a new doctrine, rather to renovate the old one.

From this perspective, the analysis conducted through the interpretative tools of CDA has proved to be extremely fruitful to test the innovative value of the Council documents and the open-mindedness of the theologians who attended the council and their careful attitude towards non-specialized readers.

"If the language of faith ceases to be in dialogue with the experience of the world, it has effectively become the language of unbelief" (Ebeling 1973, p. 192).

At this stage, we have not compared the present findings with similar sources of religious prose yet. The main aim of our future work is to carry out a comparative analysis with other documents of *Magisterium Ecclesiae* across different time periods in order to establish whether the innovative elements attested in our corpus reflect an isolated and corpus-specific linguistic phenomenon or an entire community of speakers' linguistic habit. It is furthermore worth investigating the syntactic transformations of the basic clause patterns occurring in religious prose which may contribute to a better understanding of the ideational, the interpersonal and the textual functions performed by the English language in the texts under study. Regular occurrence of the same features / innovations in other sources across time would provide evidence of their stabilization in religious prose and this will be the next task of the present ongoing research.

6. References

Aarts, B., Meyer, C.F. (eds) (1995). *The verb in contemporary English. Theory and description.* Cambridge: CUP.

Aarts, B., Close, J., Leech, G., Wallis, S. (2013). *The Verb Phrase in English. Investigating Recent Language Change with Corpora*. Cambridge: CUP.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* London: Longman Press.

Bick, E. (2000). *Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus: Aarhus Univ. Press.

Basil, C.B. (1981). *The theology of Vatican II.* London: Christian Classics.

Clark, U. (2007). *Studying Language: English in Action.* New York: Palgrave Macmillan

Crystal, D. (2010). *The Cambridge Encyclopedia of the English Language.* Cambridge: CUP.

Ebeling, G. (1973). *Introduction to a Theological Theory of Language.* London: Collins.

Faiclough, N. (2010). *Critical Discourse Analysis. The Critical Study of Language.* London: Longman.

Fowler, R. (1996). *Linguistic Criticism.* Oxford: OUP.

Karlsson, F., Voutilainen, A., Heikkilä, J. & Antilla, A. (1995). *Constraint Grammar: A Language-independent system for Parsing Unrestricted text.* Berlin: Mouton de Gruyter.

O'Collins, G., Farrugia, M. (2003). *Catholicism, The Story of Catholic Christianity*. Oxford: Oxford University Press.

Pope John XXIII, Opening of the Second Vatican Council), http://www.vatican2voice.org/91docs/convoke.htm.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*. London and New York: Longman.

Ratzinger, J. (2008). *Jesus of Nazareth: From the Baptism in the Jordan to the Transfiguration*. Rome: Ignatius Press.

Rogers, R. (2004). *An Introduction to Critical Discourse Analysis in Education*. Mahwah, NJ: Lawrence Erlbaum.

Simpson, P. (2004). *Stylistics. A resource book for students.* London and New York: Routledge.

Tanner, N.P. (1988). *The Church in Council. Conciliar Movements, Religious Practice and the Papacy from Nicea to Vatican II.* London: I.B.Tauris & Co Ltd.

Tanner, N.P. (1990). *The Decrees of the Ecumenical Councils.* Washington: Georgetown University Press.

Walter, K. (1989). *Theology and Church.* London: SCM Press.

Humour and non-humour in religious discourse

Daniela Gîfu[1], Liviu-Andrei Scutelnicu[1,2], Dan Cristea[1,2]

[1]„Alexandru Ioan Cuza" University, Faculty of Computer Science, 16,
General Berthelot, 700483, Iași

[2]Institute of Computer Science, Romanian Academy - Iaşi branch,
„T. Codrescu", 700481, Iași

E-mail: {daniela.gifu, liviu.scutelnicu, dcristea}@info.uaic.ro

Abstract

The paper describes a pilot study focusing on the exploration of humorous features in religious texts. Although religious discourses have a strong dogmatic tonality, some constructions contain visible humorous features, especially adapted to the audience expectations. Humour facilitates the apprehension of the religious discourse, evidencing not only the oratory dimension of the speaker but also helping the receptor to better perceive the message while also inducing affective effects. A corpus of preaches (which is contrasting with liturgical texts) is collected, in which humour is marked on Adjectival Noun Phrases. We propose a patter-based method for identifying humour in religious discourses in which patterns are lexicalised regular expressions of word categories. Using a religious lexicon, we classified Adjectival Noun Phrases in religious and non-religious. The study is meant to create a tool for automatic detection of humour in religious discourses. Automatically annotated corpora of preaches could become sources for further research that would reveal different valences in the religious texts.

Keywords: detection of humour, religious lexicon, Adjectival Noun Phrases.

1. Introduction

The motivation for our study relies on the need for objectivity in the interpretation of the humour in religious language. In this sense, we will explore the applicability of computational approaches to the recognition of adjectivally expressed humour. Our goal is to recognize non-verbal humour (here, in Adjectival Noun Phrases) as other research on this issue so far focused mostly on the recognition of verbal humour (Loehr, 1996; Wolfe, 2011, Mihalcea, 2012).

Marked by a strong inter-textuality (Zafiu, 2010), involving its own discursive practices, religious speech (here, preaches) requires special attention to the construction and adaptation in order to make it easier to understand by people of a very heterogeneous cultural lead (Dincă, 2008). Humorous statements can produce unexpected, sometimes contrary, effects in the hearer – cohesion (Săftoiu, 2006; Constantinescu, 2006) and exclusion (Koestler, 1964) – and, as such, an objective evaluation becomes essential. We note that humour is a form of communicative behavior for which the transmitter expects an immediate reaction (a certain type of emotion) in the receiver. Humour becomes ironic (Reboul, 1980) when the playful is substituted by offensive intentions and the irony manifestation attracts some limitations (Sălăvăstru, 1995).

Viewed as a set of discursive actions with a religious specificity, religious discourse, with its forms – preach, homily, panegyric, paraenesis, religious conference (Gordon, 2001; Grigoraş, 2000) – can be defined by the intention for producing statements which induce a certain type of emotion in the receiver.

In this study we concentrate on preach, considered to be the most relevant type of discourse in the inherited tradition of Aristotelian rhetoric (Aristotel, 2004) in the European culture and, consequently, in the Romanian culture. A preach becomes persuasive by making use of three components, each of them having specific frequencies: the ethos (speaker individuality), the pathos (use of emotion), the logos (use of rational arguments). Humorous statements investigation identified in the proposed study leads to the definition of new religious contemporary oratory features. Using humour as a mark of oral speech helps perception, understanding and deepening religious message, contribute to group cohesion, moving the rapprochement dominance between the transmitter and the auditor.

The following forms of humour are often present in preaches.

1. Irony is that type of humour through which the speaker expresses the opposite of what the audience expects, implying the words that reflect the discrepancy between appearance and essence. Here is an example:

Sunt unii care îmi spun să mă rog pentru ei și, când mă duc la Domnul, să nu-i uit. Da! Frăția ta mănâncă și dormi până te saturi, și eu am să mă rog pentru tine![1].

The speaker mocks the receiver naivety in carrying out Christian's duties, reminding him that he does violence to the Christian morality values. In some situation, the irony has to penalize the humans' ignorance and to make people aware of the ethical values.

[1] (EN) - There are people that ask me to pray for them when I go to God, not to forget them. Yes! My brother, eat and sleep till you are fulfilled and I will pray for you!

2. Frames' humour can be distinguished at the level of serious tonality and at the humorist level of the speech. The joke is based on common knowledge and values, making the audience feel comfortable.

For instance:

Dacă vrei să mă fotografiezi, caută un măgar, fă-i poză și scrie pe ea Cleopa.[2]

The donkey is a burden animal and, similarly, the speaker feels saddled with the desire of people to photograph him. In addition, the donkey has big ears and looks funny. The preacher chooses this form of expression in order to put the photographers to shame and to show them the ridiculous situation of the speaker would have when taking photos.

3. Wordplay is achieved by joining disjoint meanings of words, see the example:

Răbdare, răbdare, răbdare.... Nu până la prășit, că nu spune Sfînta Evanghelie așa. Cine va răbda până la sfârșit, acela se mântuiește, ci nu până la prășit.[3]

The speaker makes indirect reference to the way the patience is perceived in a laic sense (temporarily), in total opposition to the biblical sense (for good). The substitution of the word *sfârșit* with *prășit* induces a humorous tone. It is an inspired wordplay that has rhyme and rhythm, both words having two syllables and the same ending – easy to remember.

In this paper we propose a new method of detecting and recognizing the humour in religious discourses. In particular, we investigate whether semi-automatic classification techniques can be seen as a viable approach to distinguish humorous sequences between religious and non-religious texts. We demonstrate with empirical evidence that the humour in religious texts can be detected by automatic means.

The paper is structured as follows. Section 2 shortly describes the background. Section 3 describes the resources used in this paper. Section 4 discusses the methodology applied in the recognition of adjectival humour in religious texts. Section 5 presents statistics and their interpretation. Finally, Section 6 depicts some conclusions and directions for future work.

2. Background

Analysed by humanists (linguists (Attardo & Raskin, 1994) and psychologists (Freud, 1928; Ruch, 2002)), humour has recently raised the interest of computer scientists, who are concerned with the construction of language models for the automatic recognition and categorization of the humorous style (Stock & Strapparava, 2003), or to suggest relevant indicators of humour, authors as Ruch, Bucaria (Ruch, 2002; Bucaria, 2004) focused on alliteration, antonym, and adult slang.

Based on the narrative strategies adopted by the orator, different types of humour often found in discourse, such as anecdote, exaggeration, irony, satire, underestimation, humorous situation, (Solomovici, 2002) are mentioned as sarcasm, exaggeration or minimization, self deprecation, teasing, rhetorical question answers, double meanings, punning, proverbs interpretation (Martin, 2007).

Humour provides oratory depth, even to the religious language, accomplishing also an important social role. The humour theories are seen as complementary (Raskin, 1998; Rutter, 1997; Minsky, 1981). While it is merely considered a way to induce hilarity, humour can have positive effects: it alters attention and memory (Baym, 1995); facilitates social interactions, helping to generate solidarity and group identity (Binsted & Ritchie, 1997; Binsted, et al. 2006); improves communication problems (Bergson, 1980); can establish a common ground between dialogue partners (Hewitt, 2002); enhances the motivation, attention, understanding and capturing of information and gives an affective meaning to the message by bringing into scene an affective sense (Nijholt, 2006). As a primary mechanism for establishing the individuality of humans, humour makes the speaker feel appreciated when the receivers recognize her/his jokes and this improves communication (Black & Forro, 1999), stimulates creativity, memory, and improves the morale and the productivity of speeches (Stock & Strapparava, 2003). But humour can also have negative influences: it may offend, can inhibit the communication when jokes are too harsh, or can create professional stress (Black & Forro, 1999).

In the papers mentioned, we observed that few attempts have been made to develop systems for automatic humour recognition (Mihalcea, 2012). This is obvious, since, from a computational perspective, humour recognition appears to be significantly more subtle and in consequently, difficult to be examined.

In this work, we explore the applicability of computational methods to the recognition of Adjectival Noun Phrases (ANPs) expressed humour in religious speech. Moreover, we investigate whether automatic classification techniques represent a viable approach to distinguish between humorous and non-humorous text especially in preaches.

3. Resources and pre-processing

The study of religious language should necessarily be approached in an interdisciplinary way, in which the rhetoric sciences, communication and doxology cooperate with computational linguistic methods.

To prepare the corpus of humour in religious texts we have processed a collection of texts summing up 16 volumes of preaches in Romanian, authored by the monk priest Ilie Cleopa[4], one of the most renowned Romanian Orthodox orators and religious writers. The collection of

[2] (EN) - If you want to take me a photo, look for a donkey, take him a picture and write on it Cleopa.

[3] (EN) - Have patience, patience, patience... not till hoeing, because The Holly Gospel does not say so. But he who endures to the end will be saved, but not till the time of hoeing.

[4] Cleopa, I. (1995-1999). Ne vorbește Părintele Cleopa, vol. 1-16, Ed. Episcopiei Romånului și Hușilor.

texts, containing many humorist sequences, was published between 1995 and 1998 by one of his closest disciples. The texts count 588.784 words, over approximately 1,500 pages. The corpus was pre-processed by tokenising it, then tagging it at part-of-speech (POS), with the Romanian POS-tagger web-service (Simionescu, 2011), lemmatising it and then extracting Adjectival Noun Phrases. We considered the adjectives to be semantically relevant for the religious genre.

A lexicon of religious terms was developed using as seeds the lexicon of the semantic class RELIGIOUS used in the research on political discourse analysis in election (Gîfu & Cristea, 2012). The RELIGIOUS class is one of the 30 semantic classes, which are considered to optimally cover the necessity of interpreting the political discourse in electoral contexts, in the Discourse Analysis Tool (Gîfu & Cristea, 2012). The hierarchy of these categories preserves the structure of a tree. Then, this lexicon was enriched by further importing synonyms from DEX-online[5], the greatest public online dictionary for Romanian. DEX has no semantic structure (unlike WordNet, for instance). With this lexicon under eyes, one of the authors went through a tedious process of manual annotation of the corpus for humorous religious and non-religious ANPs. In most of the cases these expressions are a combination of religious word + non-religious word, with a humorous slant: duh necurat[6]; dracul milostiv[7]; *credință strâmbă*[8].

Finally, the lexicon was completed with hyponyms of the Religion synset in the Romanian WordNet (Tufiș et al., 2004)[9]. When this process was finalised, the religious lexicon contained 2367 entries, considered to cover satisfactorily this type of analysis.

4. The methodology

The research followed the following steps: a). pre-processing the Corpus; b). manual annotation of the humorous Adjectival Noun Phrases in the Corpus; c). automatic detection of humorous ANPs; d). evaluation.

We will call a sequence of one or more adjectives connected by conjunctions and/or commas and modifying a noun an Adjectival Noun Phrase. Whether the semantic category of the content words (nouns and adjectives) is religious or not, we will have Religious Adjectival Noun Phrases (RANP) and Non-Religious Adjectival Noun Phrases (NRANP). These phrases have been extracted by applying lexical-syntactic patterns within the borders of noun phrases. Examples are given below.

In standard Romanian, usually adjectives stay after nouns, and our religious texts statistically follow confidently this rule (for instance, pomi neroditori or *atenție dumnezeiască*[10]). Examples of adjectival constructions follows:

noun + article + adjective - Origen cel blestemat[11]:

```
<ANP id="8.13" type="non-religious">
  <W EXTRA="NotInDict" LEMMA="Origen"
MSD="Np" POS="NOUN" Type="proper"
id="150.15" offset="93">Origen</W>
  <W Case="direct" Gender="masculine"
LEMMA="cel" MSD="Tdmsr" Number="singular"
POS="ARTICLE" Type="demonstrative"
id="150.16" offset="100">cel</W>
  <W Case="direct" Definiteness="no"
EXTRA="Participle
Lemma:blestema(tranzitiv)″
Gender="masculine" LEMMA="blestemat"
MSD="Afpmsrn" Number="singular"
POS="ADJECTIVE" id="150.17"
offset="104">blestemat</W>
</ANP>
```

noun + adjective + conjunction + adjective: *cărturarilor nebuni și orbi*[12]:

```
<ANP id="8.30" type="non-religious">
  <W Case="oblique" Definiteness="yes"
Gender="masculine" LEMMA="cărturar"
MSD="Ncmpoy" Number="plural" POS="NOUN"
Type="common" id="443.4"
offset="10">cărturarilor</W>
  <W Case="direct" Definiteness="no"
Gender="masculine" LEMMA="nebun"
MSD="Afpmprn" Number="plural"
POS="ADJECTIVE" id="443.5"
offset="23">nebuni</W>
  <W LEMMA="și" MSD="Cc" POS="CONJUNCTION"
Type="coordinating" id="443.6"
offset="30">și</W>
  <W Case="direct" Definiteness="no"
Gender="masculine" LEMMA="orb"
MSD="Afpmprn" Number="plural"
POS="ADJECTIVE" id="443.7"
offset="33">orbi</W>
</ANP>
```

noun + (adjective + comma)* + adjective + conjunction + adjective: *oameni aleși, drepți și sfinți*[13]:

```
<ANP id="8.23" type="religious">
  <W Case="direct" Definiteness="no"
Gender="masculine" LEMMA="om" MSD="Ncmprn"
Number="plural" POS="NOUN" Type="common"
id="55.39" offset="205">oameni</W>
```

[5] http://dexonline.ro

[6] (EN) - unclean spirit.

[7] (EN) - the merciful devil.

[8] (EN) - crooked faith.

[9] http://www.racai.ro/

[10] (EN) - unfruitful trees or divine attention.

[11] (EN) - the cursed Origen.

[12] (EN) - to the fool and blind scribes.

[13](EN) - Chosen, honest and saint people.

```
  <W Case="direct" Definiteness="no"
EXTRA="Participle Lemma:alege(tranzitiv)"
Gender="masculine"   LEMMA="ales"
MSD="Afpmprn" Number="plural"
POS="ADJECTIVE" id="55.40"
offset="212">aleşi</W>
  <W LEMMA="," MSD="COMMA" id="55.41"
offset="217">,</W>
  <W Case="direct" Definiteness="no"
Gender="masculine" LEMMA="drept"
MSD="Afpmprn" Number="plural"
POS="ADJECTIVE" id="55.42"
offset="219">drepţi</W>
  <W LEMMA="şi" MSD="Cc" POS="CONJUNCTION"
Type="coordinating" id="55.43"
offset="226">şi</W>
  <W Case="direct" Definiteness="no"
Gender="masculine" LEMMA="sfânt"
MSD="Afpmprn" Number="plural"
POS="ADJECTIVE" id="55.44"
offset="229">sfinţi</W>
</ANP>
```

As can be seen in the above examples, the corpus includes on XML level basic level annotation that marks morpho-syntactic features and word lemmas attached to each token (<W></W>). Above this level, Adjectival Noun Phrases have been marked as XML elements <ANP></ANP>. This way we draw the attention on the adjectives that complement nouns (also including articles, conjunctions, commas and/or other linguistic connectors). The religious/non-religious semantics of the content words belonging to the ANP elements (as expressed by their inclusion in the religious lexicon) is coded in the attribute TYPE of the ANP element. A humorist tonality characterises both types of expressions.

Figure 1 shows a classification of the ANPs in the corpus. They could include or not humorous effects (HANP and NHANP), and the Humorous Adjectival Noun Phrases could be religious and non-religious (RHANP and NRHANP).

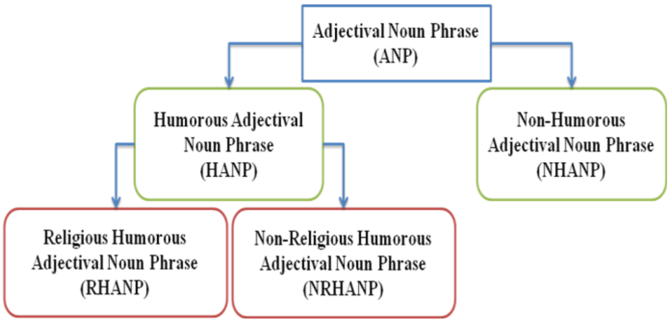


Figure 1: A classification of the Adjectival Noun Phrases

5. Statistics and interpretation

Out of the complete set of Adjectival Noun Phrases, the Gold corpus puts in evidence only those including a humorist effect, they being explicitly categorised in 2 classes: religious and non-religious. The fact that the Gold corpus lacks an explicit notation of humour within an exhaustive set of ANPs has two reasons: first, we relied on the high precision and recall of recognising ANPs (pattern-based) and to the high precision of the tags left behind by the POS-tagger (Tufiş & Dragomirescu, 2004), and second, we wanted to accelerate the process of manual annotation (this way, the annotator paid attention only to sequences displaying different kinds of humour). Identification of types of humour, although of interest, has not been considered in this research. The instances of ANPs extracted from the annotated corpus were used to form the collection of patterns of the recognition. Only variations in word forms were accepted in this phase of the research, therefore we could call our patterns lexicalised regular expressions of word categories. Table 1 compares the automatic detection of humour with that manually annotated in the Corpus with respect to religious/non-religious sequences.

Between aRHANP and mRHANP we identified 6932 common Adjectival Noun Phrases and between aNRHANP and mNRHANP we identified 1019 common Adjectival Noun Phrases.

Total Number of Words	588.784
manual Humorous Adjectival Noun Phrases (mHANP)	9937
automatic Humorous Adjectival Noun Phrases (aHANP)	9854
manual Religious Humorous Adjectival Noun Phrases (mRHANP)	8354
automatic Religious Humorous Adjectival Noun Phrases (aRHANP)	8623
manual Non-Religious Humorous Adjectival Noun Phrases (mNRHANP)	1583
automatic Non-Religious Humorous Adjectival Phrases (aNRHANP)	1231

Table 1: Automatic and manual annotation results.

With these values, the Precision (1), Recall (2) and F-measure (3) could be computed, for both Religious and Non-Religious Humorous ANPs.

$$P = \frac{\#correctly\_identified\_ANP}{\#automatically\_identified\_ANP} \quad (1)$$

$$R = \frac{\#correctly\_identified\_ANP}{\#manually\_annotated\_ANP} \quad (2)$$

$$\mathrm{F-measure} = \frac{2 * \mathrm{P} * \mathrm{R}}{\mathrm{P} + \mathrm{R}} \quad (3)$$

The values are given in Table 2.

	Precision	Recall	F-measure
RHANP	81%	83 %	82%
NRHANP	83%	64%	71%

Table 2: Statistical results for the detection algorithm

As shown in Table 2 the results for the automatic detection of the adjectival phrases, which have a religious and non-religious nature, are rather high. The fact that the religious humour is scored better than the non-religious humour, could be due to the special attention that we paid on annotating religious terms.

6. Conclusions and future work

This research is a preliminary study in humour recognition in Orthodox preaches and it confirms the hypothesis that humour, defining a specific rhetoric in religious speech, can be depicted by automatic means. However, more efforts are necessary to stabilise these preliminary results, and to look for combined methods of humour detection.

Humour analysis may require multiple interpretations. In preaches, the text approaches colloquial language, without letting aside clarity, accuracy and theological fairness. On the other hand, at the liturgical level a sacred communication act could be evidenced, in which the priest enters into dialogue with God. The preach comes to clarify the liturgical text, which is why there is a permanent adequacy audience expectations and needs. S-a trezit cu totul nebun, cu totul stricat la minte[14], *Trupul ăsta este o mână de pământ spre* mâncarea viermilor[15], or *Ai văzut vreodată vreo femeie cu două capete? Nu se poate. Bărbatul este cap şi femeia* este trup[16], are examples which display a subtle irony of the priest. These are situations for our future work. Although the adjectival phrases are absent, the irony exists in these examples, but our method is yet unable to detect it.

From the perspective of formative function, religious language becomes the key to the receivers' universe and the language facts require adaptation to specific communication situations.

The results revealed in this study may provide a basis for a new Orthodox oratory identity in the local area. Automatically annotated corpora of preaches could become sources for further research that would reveal different valences in the religious texts. For enlarging our corpus we will take into consideration other priests' speeches that manifest similar styles with Pr. Cleopa's (we think of the former Metropolitan Bartolomeu Anania of Cluj, Constantin Necula of Sibiu, Calistrat Chifan of Bârnova Monastery of Iassy).

Another challenge is the classification of adjectival phrases in subclasses that could enhance the accuracy and allow for a finer typification of the humorous sequences.

Finally, we intend also to bring in the research machine learning methods, based on neural networks that would automatically generalise patterns from the annotation. A good neural network for pattern recognition can be feed-forward, the network being trained to associate outputs with the input patterns, in our case different types of humorous Adjectival Noun Phrases.

7. References

Aristotel. (2004). Retorica, I, 2, 1356a, trad. M.-C. Andrieş, Bucharest, IRI.

Attardo, S. and Raskin, V. (1994). Non-literalness and non-bona-fide in language: An approach to formal and computational treatments of humor. Pragmatics and Cognition.

Baym, N. (1995). The performance of humor in computer-mediated communication in Journal of Computer-Mediated Communication, 1(2).

Bergson, H. (1980). Laughter: An Essay on the Meaning of the Comic in Comedy, Ed. The Johns Hopkins University Press.

Binsted, K. and Ritchie, G. (1997). Computational rules for generating punning riddles. HUMOR, the International Journal of Humor Research, Mouton de Gruyter, 10(1).

Binsted, K., Bergen, B., Coulson, S., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A. and O'Mara, D. (2006). Computational humor in IEEE Intelligent Systems, 21(2), pp. 59-69.

Black, L. and Forro, D. (1999). Humor in the academic library: You must be joking! or, how many academic librarians does it take to change a light-bulb?, Michigan State University Libraires at Michigan State University.

Bucaria, C. (2004). Lexical and syntactic ambiguity as a source of humor. Humor, 17(3).

Constantinescu, M. (2006). *Umorul conversaţional* – strategii de sprijin, in Pană Dindelegan 2006, pp. 405-412.

Dincă, G. (2008). *Strategii de construcţie a predicii*, ILB-SIL, I [Engl. Ab.].

Freud, S. (1928). Humor in International Journal of Psychoanalysis", New York, vol. 9, pp. 1-6.

Gîfu, D. and Cristea, D. (2012). Multi-dimensional analysis of political language in Future Information Technology, Application, and Service, vol. 1/164, James J. (Jong Hyuk) Park, Victor C.M. Leung, Cho-Li Wang, Taeshik Shon (editors), Springer Science+Business Media Dortdrecht, pp. 213-221.

Gordon, V. (2001). *Bunul nume al propovăduirii, condiţie sine qua non a reuşitei predicii* in *Teologie şi viaţă*, nr. 1-6, Iaşi, 104.

Gordon, V. (2000). "Introducere în omiletică", Ed. Universităţii din Bucureşti, pp. 253-254.

Grigoraş, C. (2000). *Propovăduiţi Evanghelia la toată făptura*, Iaşi, Ed. Trinitas, 29(32).

Hewitt, J. (2002). The Architecture of Thought: A New Look at Human Evolution, Holmhurst House, Beds.

Koestler, A. (1964). Act of Creation, London, Hutchinson.

Loehr, D. (1996). An Integration of a Pun Generator with a Natural Language Robot. Proceedings Twente Workshop on Language Technology 12 (TWLT 12), Computational Humor: Automatic Interpretation and Generation of Verbal Humor, Eds. J. Hulstijn and A.

[14] (EN) - He woke up completely insane, completely mind broken.

[15] EN) - This body is handful of earth that is food for worms.

[16] (EN) - Have you ever seen a woman with two heads? You cannot. The man is the head and the woman is the body.

Nijholt, pp. 161-172.

Martin R.A. (2007). The psychology of Humor, An Integrative Approach, Amsterdam: Elsevier Academic Press.

Mihalcea, R. (2012). The language of Humour in Computational Humor, Anton Nijholt (ed.), Amsterdam, pp. 5-6.

Minsky, M. (1981). Joke an their Relation to the Cognitive Unconscious in Cognitive Constraints on Communication, Reidel, Ed. Naina and Haintikka.

Nijholt, A. (2006). Conversational agents a little humor too in IEE Intelligent Systems, pp. 22-26.

Raskin, V. (1998). The sense of humor and truth in Ruch, W, The Sense of Humor: Explorations of Personality Characteristic, Berlin, Mouton de Gruyter, pp. 95-108.

Reboul, O. (1980). Langage et idéologie, Paris, Press Universitaires de France.

Ruch, W. (2002). Computers with a personality? Lessons to be learned from studies of the psychology of humor in Proceedings of The April Fools Day Workshop on Computational Humour.

Rutter, J. (1997). Stand-Up as Interaction: Performance and Audience in Comedy Venues, University of Salford, chapter 2.

Săftoiu, R. (2006). *Observaţii asupra funcţiilor râsului în conversaţia fatică* in Pană Dindelegan 2006, pp. 517-524.

Sălăvăstru, C. (1995). *Logică şi limbaj educaţional*, Bucharest, Ed. Didactică şi Pedagogică, pp. 206-210.

Simionescu, R. (2011). UAIC Romanian Part of Speech Tagger, Master's degree thesis, coord. by Prof. PhD. Dan Cristea, "Alexandru Ioan Cuza" University of Iași.

Solomovici, T. (2002). 5000 de ani de umor evreiesc - o antologie subiect*ivă*, Bucharest, Ed. Tesu.

Stock, O. and Strapparava, C. (2006). Computational humor and prospects for advertising in Rob Milne: A Tribute to a Pioneering AI Scientist, Entrepreneur and Mountaineer. IOS Press.

Stock, O. and Strapparava, C. (2003). Getting serious about the development of computational humour in Proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico.

Tufiș, D. and Ion, R. and Ide, N. (2004). Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets in Proceedings of the 20th International Conference on Computational Linguistics, COLING2004. Geneva.

Tufiș, D. and Dragomirescu, L. (2004). Tiered Tagging Revisited in Proceedings of the 4th LREC Conference, Lisabona, pp. 39-42.

Wolfe, J. (2011). Stowaway Speakers: The Diasporic Politics of Funny English in A Night at the Opera in The English Languages: History, Diaspora, Culture, vol. 2.

Zafiu, R. (2010). *Ethos, pathos şi logos în textul predicii* in Al. Gafton, Sorin Guia şi Ioan Milică, Text şi discurs religios, vol. 2, „Alexandru Ioan Cuza" University from Iasi publishing house, p. 497.

ABaC:us revisited – Extracting and Linking Lexical Data from a historical Corpus of Sacred Literature

Claudia Resch[1], Thierry Declerck[2], Barbara Krautgartner[1], Ulrike Czeitschner[1]
[1]Austrian Academy of Sciences – Institute for Corpus Linguistics and Text Technology (ICLTT)
Sonnenfelsgasse 19/8, A-1010 Vienna, Austria
[2] Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3, D-66123 Saarbrücken, Germany
E-mail: claudia.resch@oeaw.ac.at, declerck@dfki.de, barbara.krautgartner@oeaw.ac.at, ulrike.czeitschner@oeaw.ac.at

Abstract

In this submission we describe results of work within the ABaC:us project dedicated to the extraction of lexical data from a corpus of sacred literature written in historical German language: All tokens occurring in the corpus have been semi-automatically mapped onto their corresponding lemmata in modern High German, which is one of the major achievements of the project. We are currently developing a RDF and SKOS based model for the extracted lexical data, in order to support their linking to corresponding senses in the Linked Open Data (LOD) framework, with a focus on religious themes. We describe first the achieved state of the extracted lexicon and then the actual state of our semantic model for the lexicon, with some examples of the code and of the linking to lexical senses available in the LOD.

Keywords: historical language, sacred literature from the Baroque era, Linked Open Data

1. ABaC:us design

The Austrian Baroque Corpus is a growing digital collection of printed German language texts dating from the Baroque era, in particular the years from 1650 to 1750. The project group which established ABaC:us focused on the creation of a thematic research collection based on the prevalence of sacred literature during the Baroque Era. Books of religious instruction and works concerning death and dying were a focal point of Baroque culture. The best-selling genre was even marked by a number of reprinted editions. Therefore, the ABaC:us collection holds several historical religious texts specific to this genre including sermons, devotional books and works related to the dance-of-death theme:

Figure 1: Overview of the works being part of ABaC:us

In building up the corpus from scratch we avoided using already existing editions and only chose originally authentic textual data. As a matter of philological principle, we tried to get early and if possible the first known editions and rare specimens from different libraries. All Baroque prints that served as the input for the resource have been fully digitized, transcribed, and transformed in an XML format according to the guidelines of the Text Encoding Initiative (version P5).[1]

2. Annotation: PoS and lemma

ABaC:us currently contains more than 210.000 running words. All texts are enriched with different layers of information that cover basic structural annotations like chapters, headings, paragraphs and verse lines. A main part, approximately 180.000 tokens (85%), has also been tagged with part of speech (PoS) and lemma information. Thus each token was mapped automatically to a word class with the tool *TreeTagger*[2] using the 54-part so-called *Stuttgart-Tübingen-TagSet* and its guidelines (1999)[3] and was enriched with its lemma or canonical form (according to "Duden"[4] or "Deutsches Wörterbuch" by Jacob and Wilhelm Grimm[5] as a reference).

Since the automatic tagger was developed for contemporary, modern German language and was used out of domain, the project group had to cope with many erroneous mappings: The historical language used in the ABaC:us collection is a variety of Early Modern German, which at that time was not fully standardized at all. Even the slightest deviations in orthographic conventions (e.g. different suffixes, doubling of consonants, elimination of long vowels, etc.) made a historical word form difficult to classify: the verb "list" (New High German: "liest", *reads*) was automatically assigned with a noun label (NN) for "List" (*artfulness*), although it should correctly be tagged as a finite verb (VVFIN).

Because of non standard spelling the tagger cannot recognize all tokens as German words and treats them as foreign language elements (FM) or proper names (NE).

[1] See http://www.tei-c.org/Guidelines/P5/ for more details.
[2] See Schmid, 1995.
[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf
[4] See https://www.duden.de/
[5] See http://dwb.uni-trier.de/de/

As former analyses by Hinrichs and Zastrow (2012, p. 12) have shown, the tagger classifies unknown words mostly as proper names (NE). Our results confirm this observation: NE labels are mistagged with the highest frequency; only 18% of all the assigned FM elements were correct, all others were wrong.

It was a fact that the quite variable orthography caused many noisy annotations which had to be subsequently verified and manually corrected. In order to identify, to process and to remove all automatically assigned mistaggings we made use of a *token_editor* recently developed at ICLTT. This tool facilitates the evaluation of the automatic assignment of word labels, allows to verify and if necessary to correct the results by annotators. The correction of PoS and lemmas was a very time-consuming process – but after having rectified the first three texts we had a sound basis for the following ones and made our first experiments in adapting the specific historical language material for further annotation procedures.

2.1 Improvement through reliable data

For our experiments we chose two texts from the Viennese theologian and discalced Augustinian monk Abraham a Sancta Clara (1644-1709), an admired preacher and widely-read author. The writings of the preacher were very popular because of the peculiar style concerning his creation of words and other stylistic devices. His style followed the contemporary tradition of his time period. Hinrichs and Zastrow (2012) already noticed that – compared to other texts – Abraham a Sancta Clara's style exhibits by far the highest average sentence length, which might be the reason why their test data had "the highest number of tagging errors". (Hinrichs/Zastrow, p. 11)

For our purposes we have selected Abraham a Sancta Clara's bestseller "Mercks Wienn" (1679) and his subsequent book entitled "Lösch Wienn" (1680), which was also very successful. Our hypothesis was that these two works had many features in common: The texts were written by the same author, during the same time period and had a similar topic. Thus one could assume that parts of the used vocabulary would overlap.

As a consequence our intention was to create an extended lexicon with domain- and genre-specific words from "Mercks Wienn" to use its gold standard tags (approximately 57.900 tokens) as reliably tagged lexical information for tagging the unannotated text "Lösch Wienn". In order to improve the results of the ongoing process of tagging, we generated a machine-readable word list from the first text which was supposed to help to reduce the error rate of the second text:

"Lösch Wienn"	**without** lexicon of "Mercks Wienn"	**with** lexicon of "Mercks Wienn"
PoS accuracy	71,2%	82,7%
lemma accuracy	57,5%	72,4%

Figure 2: Improvement of the *TreeTagger*

A first evaluation has shown how significantly our results could be improved. The hand-corrected data[6] provided relevant lexical information to positively influence the performance of the *TreeTagger*: Not only did the PoS tagging increase in accuracy, but also the automatically assigned lemmatization worked well and exhibited a significant increase by almost 15%.

The results confirm some experiments recently carried out by Kübler and Baucom which concluded that using hand-corrected data as additional training material has a positive effect for domain adaptation. They argued "that even a fairly 'easy' problem such as PoS tagging requires a large training set" (Kübler & Baucom, 2011). Therefore we have decided to take an extended and user-defined training data set with many entries and textual material of sufficient length. The results from this experiment show that adding further entries and new lemmata to the lexicon was useful because it significantly improved the performance of the *TreeTagger*.

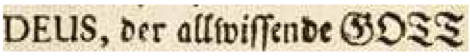
Our method of annotating more text of the same time period and genre resulted in high quality data and made ABaC:us a thoroughly proven and reliable corpus base with about 180.000 running words. In the future it will be used for:

- the annotation of more and distinct texts
- evaluating the quality of automatically generated lexical data from corpora
- the training of the *TreeTagger*
- and more sophisticated or complex linguistic research questions

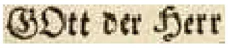
2.2 Genre specific queries

The applied linguistic annotation of ABaC:us already allows several queries that might be relevant for analyzing religious texts.

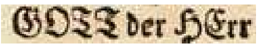
A concordance of one of the most frequent lemmata "Gott" (*god*) shows all (spelling) variants of the canonical form. Most of the occurrences are spelled with at least two capital letters, a common technique to emphasize Nomina sacra (*sacred names*):


DEUS, der allwiſſende GOTT

Deus, the omniscient Lord


GOtt der Herr

or

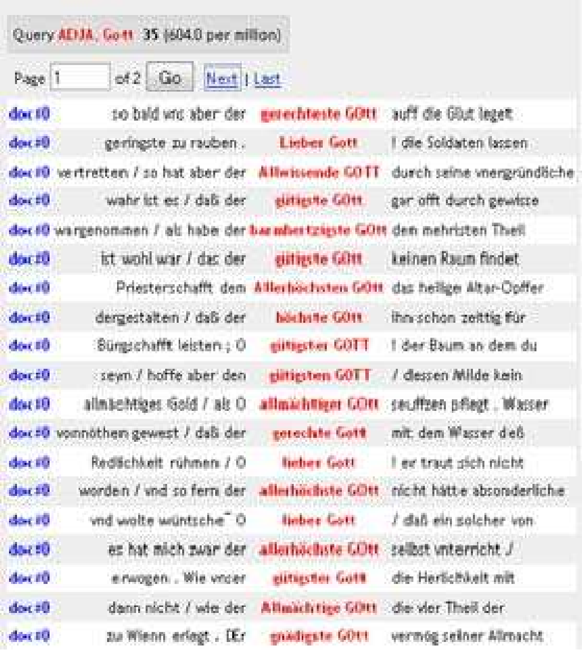

GOTT der HErr

God, the Lord

Figure 3: Examples for sacred names, taken from "Lösch Wienn", p. 48, 49 and 188.

Figure 4 (on the next page) shows the keyword in context (KWIC), as the result for the query for the lemma "Gott" with an attributive adjective as its left neighbor.

[6] Two research assistants have worked independently on the same tasks, and in case of disagreement, a senior researcher played the role of the supervisor.



Query ADJA, Gott 35 (604.0 per million)

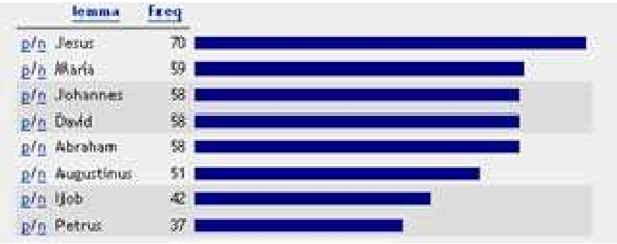
Page 1 of 2 Go Next | Last

doc #0	so bald wie aber der	gerechteste GOtt	auff die Glut leget
doc #0	geringste zu rauben .	Liebes Gott	! die Soldaten lassen
doc #0	vertretten / so hat aber der	Allwissende GOTT	durch seine vnergründliche
doc #0	wahr ist es / daß der	gütigste GOtt	gar offt durch gewisse
doc #0	wargenommen / als habe der	barmhertzigste GOtt	den mehristen Theil
doc #0	ist wohl war / das der	gütigste GOtt	keinen Raum findet
doc #0	Priesterschafft dem	Allerhöchsten GOtt	das heilige Altar-Opffer
doc #0	dergestalten / daß der	höchste GOtt	ihn schon zeittig für
doc #0	Bürgschafft leisten ; O	gütigster GOTT	! der Baum an dem du
doc #0	seyn / hoffe aber den	gütigsten GOTT	/ dessen Milde kein
doc #0	allmächtiges Gold / als O	allmächtiger GOtt	seufften pflegt . Wasser
doc #0	vonnöthen gewest / daß der	gerechte Gott	mit dem Wasser deß
doc #0	Redlichkeit rühmen / O	liebes Gott	! er traut sich nicht
doc #0	worden / vnd so fern der	allerhöchste GOtt	nicht hätte absonderliche
doc #0	vnd wolte wüntsche ̄ O	liebes Gott	/ daß ein solcher von
doc #0	es hat mich zwar der	allerhöchste GOtt	selbst vnterricht /
doc #0	erwogen . Wie vnser	gütigster Gott	die Herlichkeit mit
doc #0	dann nicht / wie der	Allmächtige GOtt	die vier Theil der
doc #0	zu Wienn erlegt . DEr	gnädigste GOtt	vermög seiner Allmacht

Figure 4: Adjectives and lemma “Gott”

Nearly all adjectives listed in Figure 4 are modified to produce a superlative form which indicates the degree of the designated property: “der gerechteste GOtt” (gerecht = *just*), “der gütigste GOtt” (gütig = *kind*), der höchste GOtt (hoch = *high*) or “DEr gnädigste Gott” (gnädig = *gracious*). The German superlative can be prefixed with “aller-” (e.g. “der allerhöchste GOtt”) and is called an excessive (Dressler & Merlini Barbaresi 1994, 558-573), meaning “the highest of all”. In this case it underlines those statements describing the omnipotence of the deity. The combination of an interjection (e.g. “o”), an adjective and the lemma “Gott” can also be found very often, particularly in prayers and phrases that express emotions and pleas.

As this kind of sacred literature contains many cross-references, quotations and names the project group decided to annotate names of historical, mythological and biblical figures (in addition to the NE tag used for proper names). The most frequent biblical names can be seen in figure 5:



	lemma	Freq
p/n	Jesus	70
p/n	Maria	59
p/n	Johannes	58
p/n	David	58
p/n	Abraham	58
p/n	Augustinus	51
p/n	Ijob	42
p/n	Petrus	37

Figure 5: Most frequent biblical names

3. Porting the extracted lexical data onto the Linked Open Data paradigm

As described in section 2, a lexicon has been semi-automatically extracted from the ABaC:us collection and manually curated. This lexical data is very valuable for further work in corpus linguistics, particularly in the field of historical texts and religious studies.

While the work presented in section 2 is mainly about establishing correspondences between historical variants of words and their actual High German lemma forms, we are also aiming at providing a semantic description to the lemmata and therefore indirectly to the form variants used in the corpus. We also aim to support wide access to the semantically enriched lexical data in a machine-readable form and investigate for this purpose the encoding of the data in the Linked (Open) Data format[7], contributing also to the emergent Linguistic Linked Data cloud[8]. Going in those directions allows us to semi-automatically link our corpus and lexical data to both domain knowledge (e.g. religion) and other language data (e.g. related lexical data within and across languages).

3.1 Pre-processing of the lexical data

In order to support our work in the field of Linked Data, we first had to re-organize the structure of the stored lexical data. The result of the work described in section 2 was stored along the lines of the findings in the corpus: one entry in the data base per occurrence of a token in the corpus. While this is essential for keeping track of the context of the word forms[9], we want to reduce the representation of the tokens to their types, as displayed in Figure 6, where the modern High German nominal lemma “Fegefeuer” (*purgatory*) is a unique entry, pointing to the list of form variants that have been detected and marked in the corpus. Therefore all variants of the modern High German lemma form “Fegefeuer” (*purgatory*) are associated – and thus identified – with this lemma. Our aim is then to link the correlated unique lemma form to available information sources, mainly in the religious domain, and so to indirectly link the tokens of the corpus to additional (semantic) information sources.

In the example in Figure 6 we also include the frequency information for each word forms in the corpus.

"Fegefeuer" => {
"NN" => {
"Feeg=Feuer" => "6"
"Feegfeuer" => "100"
"Feegfeuers" => "4"
"Fegfeuer" => "80"
"Fegfeuers" => "24"
"Fegfeur" => "2"
"Fegfewer" => "4"
}
}

[7] See http://linkeddata.org/ for more details.

[8] See http://linguistics.okfn.org/resources/llod/ for more details.

[9] Alternative, but compatible, to this form of storing the data would be a stand-off annotation schema, in which each word form (token) in the corpus is indexed.

Figure 6: Examples of older word forms associated to the modern High German lemma: “Fegefeuer”

This is the lexicon structure which we will then transform, by means of a Perl script, onto RDF-based representation languages, including SKOS-XL[10] and *lemon*[11].

3.2 RDF/SKOS/lemon representation

At ICLTT we developed a RDF-based model for representing lexical data to be used in many projects and contexts[12]. In the case of the lexical data extracted from the AbaC:us collection, we do not deal with a classical dictionary[13] as our source, but rather with a selection of word forms used in a corpus and associated with modern High German lemmata. For this reason we introduce a special owl:Class[14]:

```
    icltt:Corpus_Lexicon
        rdf:type owl:Class ;
        rdfs:comment "Lexicon extracted from a corpus"@en ;
        rdfs:label "Corpus Lexicon"@en ;
        rdfs:subClassOf owl:Thing .
```

An instance of this owl:Class is the ABaC:us data set, displayed just below:

```
icltt:abacus
    rdf:type skos:Collection , icltt:Corpus_Lexicon ;
    rdfs:label "ICLTT lexicon for Baroque language"@en ;
    skos:member icltt:concept_fegefeuer .
```

We consider such data sets as a skos:Collection rather than a skos:ConceptScheme, since we are rather listing entries than describing hierarchical or associative relations between them. We use the “skos_member” object property to mark the entries belonging to this collection, as can be seen in the example above (for lack of space we only include the entry “Fegefeuer” as a member of the collection here).

Entries are introduced at the Schema level by the owl:Class “Entry”, which is a subclass of skos:Concept:

```
    icltt:Entry
        rdf:type owl:Class ;
        rdfs:label "Entry"^^xsd:string ;
        rdfs:subClassOf skos:Concept ;
        owl:equivalentClass lemon:LexicalEntry .
```

In our model, the variants of the High German lemma forms are encoded as single lexical forms, bearing just xsd:String information. The owl:Class for this is:

```
    icltt:Form
        rdf:type owl:Class ;
        rdfs:label "Form"^^xsd:string ;
        rdfs:subClassOf skos:Concept ;
        owl:equivalentClass lemon:Form .
```

Instances of this class look like the example displayed below, introducing the language tag “fnhd” for “Frühneuhochdeutsch” (*Early New High German*):

```
    icltt:Feeg_Feuer
        rdf:type lemon:Form , icltt:Form ;
        rdfs:label "Feeg=Feuer"@fnhd ;
        lemon:formVariant icltt:Feegfeuer .
```

The corresponding instance for the High German entry:

```
icltt:concept_fegefeuer
   rdf:type lemon:LexicalEntry , icltt:Entry ;
   rdfs:label "Fegefeuer"@de ;
   lemon:lexicalForm icltt:Feegfeuer , icltt:Feeg_Feuer ;
   skosxl:prefLabel icltt:entry_Fegefeuer .
```

This instance in the last line of the code is pointing to a skos object via the property skosxl:prefLabel. We use this property to link the basic entry (as a string belonging to the corpus_lexicon) to a complex linguistic object as it is displayed below:

```
icltt:entry_Fegefeuer
   rdf:type icltt:Lemma ;
   rdfs:label "Fegefeuer"@de ;
   icltt:hasPos icltt:noun ;
   lemon:sense icltt:fegefeuer ;
   skosxl:literalForm "Fegefeuer"@de .
```

In this case the corpus_lexicon entry gets associated with PoS information, but more importantly, we add a link to a “sense”. As mentioned earlier in this paper, no “meaning” is given to us from the corpus, therefore we are querying for senses in available semantic resources in the web, more specifically in the Linked Open Data environment.

The strategy here is to send SPARQL queries to DBpedia and to see how much of our modern High German entries are present in this semantic resource. For this we use the publicly available “Virtuoso SPARQL Query Editor”, in its specialization for the German data sets.[15] Our example in this submission, “Fegefeuer”, is indeed included as a

[10] See http://www.w3.org/TR/skos-reference/skos-xl.html

[11] See McCrae & al., 2012.

[12] We use the TopBraid composer for developing our model (http://www.topquadrant.com/tools/IDE-topbraid-composer-maestro-edition/)

[13] With this, we mean that a dictionary typically lists entries of a specific language and relates them to a definition and meanings (*senses*). But the ABaC:us lexicon is closer in form to a dialectal dictionary which introduces meanings by the use of the words used in the corresponding standard language. An example of a mapping from a dialectal dictionary into SKOS is described in (Wandl-Vogt & Declerck, 2013).

[14] The examples from our ontology model are given in the turtle syntax (see http://www.w3.org/TR/turtle/ for more details).

[15] http://de.dbpedia.org/sparql

concept in DBpedia[16], and from there we acquire a lot of interesting additional information: So for example all the “redirects of”, which in this case are:

- dbpedia-de:Purgatorium
- dbpedia-de:Fegfeuer
- dbpedia-de:Reinigungsort

Of interest is the fact that there is already an alternative form included in this small list of DBpedia entries, which re-directs the string “Fegfeuer” to “Fegefeuer”. Looking again at the list of variants we have in Figure 6, we see that this word form is already included. Our task is then to include the other variant “Feegfeuer” to this DBpedia list, and so to enrich the Linked Data Framework with lexical data resulting from our corpus. Additionally, we can get further semantic information from DBpedia, such as for example the categories in which our lemmata are classified. Considering again our example “Fegefeuer”, we get these categories from DBpedia[17]:

- category-de:Eschatologie
- category-de:Mythologischer_Ort
- category-de:Tod_(Mythologie)
- category-de:Katholische_Theologie

The categories (*Eschatology, Mythological Place, Death_(Mythology) and Catholic Theology*) clearly indicate that our entry “Fegefeuer” is a relevant item in the context of religion. In this manner we are marking all lemmata in our corpus_lexicon with (religious) categories contained in DBpedia, if available.

3.3 Expressing the meaning of an entry by linking to senses in DBpedia

Now, we do not only want to add the semantic classification we can gain from resources in the LOD to our entries, but we also want to associate a lexical sense to the entries. As mentioned earlier in this submission, we do not get information about the meanings for the entries from the corpus[18], and therefore we decided to link the lexicon entries extracted from ABaC:us to senses encoded in DBpedia. As a preliminary step, we need, for the lexicon, to introduce in our model an owl:Class “Sense” for instances to which the property “lemon:sense” can point:

```
icltt:Sense
        rdf:type owl:Class ;
        rdfs:label "Sense"@en ;
        rdfs:subClassOf skos:Concept ;
        owl:equivalentClass lemon:LexicalSense .
```

In contrast to the case of entries for the lemmata, we encode the “senses” as part of a skos:conceptScheme, since in the case of senses more relations between the items are possible (and desirable):

```
icltt:Senses_ICLTT
        rdf:type skos:ConceptScheme ;
        rdfs:comment "Senses that are used in ICLTT dictionaries"@en ;
        rdfs:label "Senses"@en .
```

The instance for the sense to be associated with “Fegefeuer” is shown here;

```
icltt:fegefeuer
        rdf:type lemon:LexicalSense , icltt:Sense ;
        rdfs:label "Purgatory"@en , "Fegefeuer"@de ;
        skos:exactMatch
<http://wiktionary.dbpedia.org/page/Fegefeuer-German-Noun-1de> ;
        skos:inScheme icltt:Senses_ICLTT .
```

In this case we make use of the skos:exactMatch property to link to a sense of the LOD version of Wiktionary. One of the advantages of this approach lies in the fact that we can re-use existing semantic resources without having to create our own catalogue of senses. In addition we get a list of multilingual equivalents, as those are listed in the LOD version of Wiktionary. In the case of “Fegefeuer” we get the equivalents for English, French, Italian, Latin, Swedish, and Catalan for this one sense! And in fact, on the corresponding page for the English term[19], we get many more equivalents: approximately 50 equivalent terms in about 40 languages.

This sense-based access to lexical resources available in the LOD thus supports the creation of a multilingual network of terms relevant to religious studies. In our case, we manage to link old form variants of such religious terms and other relevant terms used in this specific religious context. For the particular example we have been discussing, we can thus not only get many multilingual equivalents for the word “Fegefeuer” (and its German variants), but also for related words that are classified under the DBpedia categories “Eschatology” etc.

4. Conclusion

We have described the actual state of a lexicon extracted from a corpus of sacred texts written in a historical German language of the Baroque era and demonstrated how it can be used in automated linguistic annotation procedures. We have shown how this lexicon, in which

[16] See http://de.dbpedia.org/page/Fegefeuer

[17] As a result of the SPARQL query:
select ?subject where
{<http://de.dbpedia.org/resource/Fegefeuer>
<http://purl.org/dc/terms/subject> $subject}

[18] But further work will be dedicated in computing meaning for word forms on the basis of their context of use, in the spirit of approaches described in the so-called distributional semantics framework (Sahlgren, 2008).

[19] http://wiktionary.dbpedia.org/page/purgatory-English-Noun-1en

nearly all tokens of the corpus are associated with modern lemma forms of German, can be used for supporting searches of the corpus, especially for topics related to religious themes. Finally we have described actual work in modeling the extracted lexicon using semantic web standards, and how this supports the linking of entries of the lexicon to lexical senses and multilingual equivalents available in the Linked Open Data framework. We will also publish some of our data in the LOD so that it can be linked to from other resources in the web of data. The lexicon will be delivered as an Austrian contribution to the CLARIN-ERIC infrastructure and we plan to make the lexicon also available in the LRE-Map.

5. Acknowledgements

We would like to thank our colleagues Eva Wohlfahrter for her assistance in annotating the ABaC:us texts and Barbara Krautgartner for creating the gold standard for our textual analyses. The current project “Text-Technological Methods for the Analysis of Austrian Baroque Literature” is supported by funds of the Österreichische Nationalbank, Anniversary Fund, project number 14738.

6. References

Bradley, J. and Pasin, M. (2012). *Annotation and Ontology in most Humanities research: accommodating a more informal interpretation context.* In: DH2012 NeDiMaH Ontology Workshop.

Czeitschner, U., Declerck, T. and Resch, C. (2013). *Porting Elements of the Austrian Baroque Corpus onto the Linguistic Linked Open Data Format.* In: Petya Osenova, Kiril Simov, Georgi Georgiev, Preslav Nakov (eds.): *Proceedings of the joint NLP&LOD and SWAIE Workshops Hissar, Bulgaria, RANLP,* pp. 12--16.

Declerck, T., Czeitschner, U., Mörth, K., Resch, C. and Budin, G. (2011). *A Text Technology Infrastructure for Annotating Corpora in the eHumanities.* In: *Proceedings of the International Conference on Theory and Practice of Digital Libraries.* In: TPDL 2011, 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 Conference on Natural Language Processing (KONVENS-10). Saarbrücken: pp. 117--121.

Dressler, W. U., Merlini Barbaresi, L. (1994). *Morphopragmatics. Diminutives and Intensifiers in Italian, German, and Other Languages*. Berlin / New York, Mouton de Gruyter.

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

Knittel, A. P. (Ed.) (2012). *Abraham a Sancta Clara. Vom barocken Kanzelstar zum populären Schriftsteller.* In: *Beiträge des Kreenheinstetter Symposions anlässlich seines 300. Todestages*. Eggingen, Edition Isele.

Kübler, S. and Baucom, E. (2011). *Fast Domain Adaptation for Part of Speech Tagging for Dialogues.* In: Proceedings of the International Conference on Recent Advances in NLP (RANLP). Hissar, Bulgaria.

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P. Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T. (2012). *Interchanging lexical resources on the Semantic Web.* In: Language Resources and Evaluation. Vol. 46, Issue 4, Springer: pp. 701--719.

Mörth, K., Resch, C., Declerck, D. and Czeitschner, U. (2012). *Linguistic and Semantic Annotation in Religious Memento Mori Literature*. In: Proceedings of the LREC'2012 Workshop: Language Resources and Evaluation for Religious Texts (LRE-Rel-12). ELRA: pp. 49--52.

Hinrichs, E. and Zastrow, T. (2012). *Linguistic Annotations for a Diachronic Corpus of German*. In: Linguistic Issues in Language Technology, Volume 7, issue 7, pp. 1--16.

Sahlgren, M. (2008). *The Distributional Hypothesis. From context to meaning: Distributional models of the lexicon in linguistics and cognitive science* (Special issue of the Italian Journal of Linguistics), Rivista di Linguistica, volume 20, numero 1, 2008.

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

Schiller, A., Teufel, S., Thielen, C. (1995). *Guidelines für das Tagging deutscher Textcorpora mit STTS.* http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf

Schmid, H. (1995). *Improvements in Part-of-Speech Tagging with an Application to German.* Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Wandl-Vogt, E. and Declerck, T. (2013). *Mapping a traditional Dialectal Dictionary with Linked Open Data*. In: Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.): *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 460--471.

Automatically generated, phonemic Arabic-IPA pronunciation tiers for the *Boundary Annotated Qur'an Dataset for Machine Learning* (version 2.0)

Majdi Sawalha[1,2], Claire Brierley[2], Eric Atwell[2]
University of Jordan[1] and University of Leeds[2]
[1] Computer Information Systems Dept., King Abdullah II School for IT, University of Jordan, Amman 11942, Jordan
[2] School of Computing, University of Leeds, LS2 9JT, UK
E-mail: sawalha.majdi@gmail.com, C.Brierley@leeds.ac.uk, E.S.Atwell@leeds.ac.uk

Abstract

In this paper, we augment the *Boundary Annotated Qur'an* dataset published at LREC 2012 (Brierley *et al* 2012; Sawalha *et al* 2012a) with automatically generated phonemic transcriptions of Arabic words. We have developed and evaluated a comprehensive grapheme-phoneme mapping from Standard Arabic > IPA (Brierley *et al* under review), and implemented the mapping in Arabic transcription technology which achieves 100% accuracy as measured against two gold standards: one for Qur'anic or Classical Arabic, and one for Modern Standard Arabic (Sawalha *et al* [1]). Our mapping algorithm has also been used to generate a pronunciation guide for a subset of Qur'anic words with heightened prosody (Brierley *et al* 2014). This is funded research under the EPSRC "*Working Together*" theme.

Keywords: IPA phonemic transcription; SALMA Tagger; Arabic transcription technology

1. Introduction: the Boundary Annotated Qur'an dataset for machine learning

For LREC 2012, we reported on a Qur'an dataset for Arabic speech and language processing, with multiple annotation tiers stored in tab-separated format for speedy text extraction and ease of use (Brierley *et al* 2012). One novelty of this dataset is Arabic words mapped to a prosodic boundary scheme derived from traditional *tajwīd* (recitation) mark-up in the Qur'an as well as syntactic categories. Thus we used our dataset for experiments in Arabic phrase break prediction: a classification task that simulates human chunking strategies by assigning prosodic-syntactic boundaries (phrase breaks) to unseen text (Sawalha *et al* 2012a; 2012b). In this paper, we report on version 2.0 of this dataset with 4 new prosodic and linguistic annotation tiers. It features novel, fully automated transcriptions of each Arabic word using the International Phonetic Alphabet (IPA), with an IPA > Arabic mapping scheme based on Quranic recitation, traditional Arabic linguistics, and modern phonetics (Brierley *et al* under rveiew).

2. The *Boundary Annotated Qur'an* for Arabic phrase break prediction

Phrase break prediction is a classification task in supervised machine learning, where the classifier is trained on a substantive sample of "gold-standard" boundary-annotated text, and tested on a smaller, unseen sample from the same source *minus* the boundary annotations. The task equates to classifying words, or the junctures (*i.e.* whitespaces) between them, via a finite set of category labels: in most cases, a binary set of `breaks` versus `non-breaks`; and less commonly, a tripartite set of `{major; minor; none}`.

2.1 Boundary annotations

A boundary-annotated and part-of-speech tagged corpus is a prerequisite for developing phrase break classifiers. One novelty of our dataset is that we derived a coarse-grained, prosodic-syntactic boundary annotation scheme for Arabic from traditional recitation mark-up, known as *tajwīd*. Tajwīd boundary annotations are very fine-grained, delineating eight different boundary types, namely: three major boundary types, four minor boundary types, and one *prohibited* stop. For our initial experimental purposes, we collapsed these eight degrees of boundary strength into the familiar `{major, minor, none}` sets of British and American English speech corpora (Taylor and Knowles 1988; Beckman and Hirschberg 1994). Another novelty is that we used certain stop types (*i.e.* compulsory, recommended, and prohibited stops) to segment the text into 8230 sentences.

2.2 Syntactic annotations: the efficacy of traditional Arabic syntactic category labels

Phrase break prediction assumes part-of-speech (PoS) tagged input text as well as prior sentence segmentation, since syntax and punctuation are traditionally used as classificatory features. Traditional Arabic grammar (Wright, 1996; Ryding, 2005; Al-Ghalayyni, 2005) classifies words in terms of just three syntactic categories {`nouns; verbs; particles`}, and another novelty of our dataset is that we retained this traditional feature set as the default. Subsequently, we added a second syntactic annotation tier differentiating a limited selection of ten subcategories extracted from fully parsed sections of an early version of the *Quranic Arabic Corpus* (Dukes 2010). These comprise `{nouns; pronouns; nominals; adverbs; verbs; prepositions; 'lām prefixes; conjunctions; particles; disconnected letters}`. However, our preliminary experiments with a trigram tagger for Arabic phrase break prediction report a significant improvement of 88.69% in respect of baseline accuracy (85.56%), using the traditional, tripartite, syntactic feature set (Sawalha *et al* 2012a; Sawalha *et al* 2012b). A sample from version 1.0 of the multi-tiered *Boundary Annotated Qur'an* dataset is shown in Fig 1. The 8 columns of Fig 1 are: (1) Arabic words in Othmani script, (2) Arabic words in MSA script, (3) three POS tags of the word, (4) ten POS tags of the word, (5) verse ending symbol, (6) tripartite boundary annotation tag, (7) binary boundary annotation tag, and (8) sentence terminal.

OTH	**MSA**	**Syntax: NVP**	**Syntax: 10 PoS**	**Verse ends**	**Boundaries: tripartite**	**Boundaries: binary**	**Sentences**
بِسۡمِ	بِسْمِ	N	NOUN	-	-	non-break	-
ٱللَّهِ	اللَّهِ	N	NOUN	-	-	non-break	-
ٱلرَّحۡمَٰنِ	الرَّحْمَٰنِ	N	NOMINAL	-	-	non-break	-
ٱلرَّحِيمِ	الرَّحِيمِ	N	NOMINAL	۝	||	break	terminal
ٱلۡحَمۡدُ	الْحَمْدُ	N	NOUN	-	-	non-break	-
لِلَّهِ	لِلَّهِ	N	NOUN	-	-	non-break	-
رَبِّ	رَبِّ	N	NOUN	-	-	non-break	-
ٱلۡعَٰلَمِينَ	الْعَالَمِينَ	N	NOUN	۝	||	break	terminal

Figure 1: Sample tiers from the *Boundary-Annotated Qur'an* dataset version 1.0

3. IPA transcription tiers for the *Boundary Annotated Qur'an* (version 2.0): rationale

One of our objectives in the EPSRC-funded *Working Together* project is to automate Arabic transcription using a carefully defined subset of the International Phonetic Alphabet (IPA). This task matches the definition of *transcription*, as opposed to *transliteration* and *romanisation* for Arabic, as described in Habash *et al* (2007). The output of our algorithm is a phonemic pronunciation for each Arabic word as an element of its citation form, similar to entries in the OALD[1] and LDOCE[2] for English, to enhance Arabic dictionaries, to facilitate Arabic language learning, and for Arabic natural language engineering applications involving speech recognition and speech synthesis.

A corpus of fully vowelised Arabic text was essential for developing and evaluating both the comprehensive grapheme-phoneme mapping from Standard Arabic > IPA, and the mapping algorithm itself. The Qur'an is an iconic text and an excellent gold standard for modeling and evaluating Arabic NLP, since it arguably subsumes other forms of Arabic, including regional dialects (Harrag and Mohamadi 2007), and MSA (Sharaf 2012). Hence we have developed our mapping and our mapping algorithm on the *Boundary-Annotated Qur'an* dataset, which includes the entire text of the *Qur'an* in fully vowelised MSA as well as traditional *Othmani* script.

A further research objective in *Working Together* is stylistic and stylometric analysis of the *Qur'an*, and a phonemic representation of the entire text via our MSA > IPA mapping will facilitate this. Version 2.0 of the *Boundary Annotated Qur'an* dataset has emerged as a result, featuring Arabic words tagged with two alternative pausal, phonemic transcriptions in IPA (one with and one without short vowel case endings), plus a Buckwalter-style transliteration and an Arabic root where applicable.

4. The Arabic > IPA mapping: linguistic underpinning

In general, Arabic spelling is a phonemic system with one-to-one letter to sound correspondence. Nevertheless, our mapping scheme is original due to its treatment of certain character sequences as *compounds* requiring transcription. This differentiates our scheme from the machine readable Speech Assessment Methods Phonetic Alphabet or SAMPA for Arabic (Wells 2002), where many more hand-crafted rules would need to be developed before implementing automatic Arabic > SAMPA transcription due to the sparseness of the scheme itself. Therefore, as well as the usual transcription of consonants, long and short vowels, and diacritic marks, we have compiled a dictionary of mapped MSA > IPA pairings that both anticipates and documents grapheme-phoneme relationships extending beyond a single letter to the immediate right-left context in fully vowelised Arabic text. For example, Arabic has two diphthongs which are each realised orthographically via the trigram sequence VCV, where V represents a short vowel or other diacritic mark and C is a consonant or semi-vowel (Fig. 2).

Arabic	Example	N-gram capture	IPA
◌َيْ	بَيْت	trigram: VCV	/aj/
◌َوْ	حَوْل	trigram: VCV	/aw/

Figure 2: Two diphthongs represented by a trigram character sequence

Our mapping is also original in that it draws on traditional Arabic linguistics for selecting the most appropriate subset of IPA symbols to represent the sound system of the language. A basic version of our Arabic > IPA mapping appears in Appendix I; the full version appears in Brierley *et al* (under review).

[1] Oxford Advanced Learner's Dictionary
[2] Longman Dictionary of Contemporary English

5. The Arabic > IPA mapping algorithm

The Arabic > IPA mapping algorithm automates phonetic transcription of Arabic words and outputs a phonemic pronunciation for each word. The algorithm has two stages: the pre-processing stage where Arabic word letters are mapped to their IPA character equivalent on a one-to-one basis; and a second stage which involves the development and application of phonetic rules that modify the IPA string produced in the first stage to produce the correct IPA transcription of the input Arabic word. The following subsections briefly describe the stages of the algorithm. The algorithm is explained in detail in (Sawalha *et al*, 2014a).

5.1 Pre-processing stage

During this project, a carefully defined subset of the International Phonetic Alphabet (IPA) for Arabic transcription was defined (Brierley *et al* under review). This includes mapping Arabic consonant letters into one IPA alphabet such as (ب، ت، ث، ...) > (/b/, /t/, /θ/), or into two IPA alphabets such as (ص، ض، ط، ...) > (/sˤ/, /dˤ/,/tˤ/). IPA alphabets for both long and short vowels are also defined, long vowels such as (ا، و، ي)> (/aː/,/uː/,/iː/), and short vowels such as (◌َ، ُ، ِ)>(/a/,/u/,/i/). *hamza*^{h} (ء، أ، ؤ، ئ), regardless of form or shape, is represented by the IPA character (ʔ). IPA alphabets for *tanwīn* are defined such that (◌ً، ◌ٌ، ◌ٍ) > (/an/, /un/, /in/); *sukūn* is not mapped to any IPA character because *sukūn* represents silence. Using these carefully defined sets of IPA alphabets, a 58-entry dictionary was constructed to facilitate the Arabic > IPA one-to-one mapping. Appendix I shows a basic version of our Arabic > IPA mapping; the full version appears in Brierley *et al* (under review).

The tokenization module of the SALMA-Tagger (Sawalha, 2011; Sawalha and Atwell, 2010) was used to tokenize and preprocess the input Arabic text. The SALMA-Tokenizer preprocesses Arabic words by resolving gemination marked by (◌ّ) *šadda*^{h} into two similar letters: the first carries a *sukūn* diacritic and the second carries a short vowel similar to the short vowel of the original geminated letter. The tokenizer also replaces the prolongation letter (آ) *madd* into *hamza*^{h} followed by the long vowel *'alif*. The SALMA-Tokenizer has a spell checking and correction module which verifies the spelling of the Arabic word in terms of valid letter and diacritic combinations. It limits each letter of the processed word to only one diacritic. The output of the SALMA-Tokenizer is an Arabic word string which best suits the one-to-one mapping of Arabic letters into IPA alphabets.

As a first step in the mapping process, the one-to-one mapping module reads the processed Arabic word. For each letter it searches the dictionary for its equivalent IPA alphabet. The output is an IPA string representing the sequence of IPA alphabets equivalent to the Arabic letters and diacritics of the input word. For example the word يَتَسَاءَلُونَ *yatasā'alūna* "they are asking one another" is mapped into the IPA string /jatasaaːʔaluwna/. The accuracy of the preprocessing stage of the Arabic > IPA mapping algorithm showed that about 70% of Arabic words in the test sample were not mapped correctly. Therefore, mapping Arabic words into IPA using one-to-one mapping only is not accurate and a rendering stage of pronunciation is needed. The following subsection discusses the rule-based stage of the Arabic > IPA algorithm that renders the produced string and generates 100% accurate results.

5.2 Rule development

As shown in the previous section, a pronunciation rendering stage is needed to produce correct phonetic transcription of Arabic words. Traditional Arabic orthography includes silent letters, and ambiguous letters such as the letters (ا، و، ي) *ʾalif*, *wāw*, and *yāʾ* which can be consonants, semi-vowels or long vowels. Also, short vowels and diacritics necessary to convey the pronunciation reliably are usually absent. Some letters appear in the orthographic word but are not pronounced and some sounds are not presented in the orthographic word altogether. The major challenges for the one-to-one mapping step are: dealing with the (i) definite article (i.e. whether the *l* is pronounced or assimilated to the following sound becoming a geminate of it), (ii) long vowels when they are pronounced as vowels, (iii) *ʾalif* of the group (ألف التفريق) which is not pronounced, (iv) words with special pronunciations, (v) *hamza*^{tu} *al-waṣl* and (vi) *tanwīn*.

The second stage of the Arabic > IPA mapping algorithm is based on especially developed rules and regular expressions to deal with cases for which the one-to-one mapping fails to generate a correct phonetic transcription. Output from the previous step was studied for the purpose of finding patterns in the mistaken transcriptions. Around 50 rules were developed and ordered correctly so that algorithm could generate the correct IPA transcriptions of input Arabic words. For example, words ending in *tanwīn al-fatḥ* which were transcribed into /aaːan/ in the first stage, are rendered by the IPA string as /an/ as in the word مِهَادًا *mihādan* "resting place" which is transcribed into /mihaːdan/. Other rules deal with the definite article (ال) when followed by a letter corresponding to coronal and non-coronal sounds[3]. If the definite article is followed by coronal sound then the IPA string /aːl/ representing the one-to-one mapping is replaced by /ʔa/ followed by a doubling of the coronal sound such as transcribing the word النَّبَإِ *an-naba'i* "news" as /ʔannabaʔi/. If it is followed by a non-coronal sound it is replaced by /ʔal/,

[3] Coronal consonants are consonants articulated with the flexible front part of the tongue. They are also known as solar or sun letters. Non-coronal consonants are known as lunar or moon letters.

such as the word الْحَقُّ *al-ḥaqqu* "the truth" transcribed in the IPA string as /ʔalħaqqu/. This process is non-trivial: Sawalha *et al* (2014a) has a detailed description of patterns of mistaken transcription and rules for correcting the output.

6. Evaluation

Evaluation of the *Boundary Annotated Qur'an* (version 2.0) focused on assessing the automatically generated Arabic > IPA transcription tiers of the BAQ corpus. The evaluation is performed using two methods: (i) measuring the accuracy of the algorithm by comparing the results against an especially constructed gold standard for evaluation; and (ii) generating a frequency list of Qur'an with automatically generated IPA transcriptions of the Qur'an word types and verifying these transcriptions by linguists specialized in tajwīd and phonology.

6.1 The Qur'an gold standard

A gold standard for evaluating the Arabic > IPA mapping algorithm was especially constructed. The gold standard consists of about 1000 words from the Qur'an, chapters 78 to 85. For each word in the gold standard, the IPA transcription was manually generated. Figure 3 shows a sample of the gold standard for evaluation. Evaluation of the output of the Arabic > IPA mapping algorithm showed an accuracy of 100%, in indication of what we had aspired for.

Chapter #	Verse #	Word #	Word	Pausal mapping - with case ending
78	1	1	عَمَّ	/ʕamma/
78	1	2	يَتَسَاءَلُونَ	/jatasaːʔaluːna/
78	2	1	عَنِ	/ʕani/
78	2	2	النَّبَإِ	/ʔannabaʔi/
78	2	3	الْعَظِيمِ	/ʔalʕaðˤijmi/

Figure 3: A sample of the Qur'an gold standard

6.2 Generation and verification of the IPA transcriptions of the Qur'an

The second method for evaluating the *Boundary Annotated Qur'an* (version 2.0) was to manually check and verify the automatically-generated IPA transcriptions of all words in the BAQ corpus. The BAQ corpus consists of 77,430 words and 17,606 word types. To reduce the time and effort of manual verification of the IPA transcription, word types were verified rather than words. A frequency list of the Qur'an was generated first, and then the IPA transcription for each word type of the BAQ corpus was verified by linguists who are specialized in both tajwīd and phonetics. This evaluation method is suitable for verifying the IPA transcriptions of the BAQ corpus words in their pausal forms while preserving case endings in the transcription. A sample of the frequency list for the first 50 word types is in Appendix II.

After manual verification of the Qur'an frequency list, we computed the accuracy of the Arabic>IPA transcription algorithm using the verified frequency list of the Qur'an. Only 91 errors in transcription were found in the Qur'an frequency list. These are 91 word types from a total of 17,606 word types in the list. Therefore, the accuracy of the automatic Arabic > IPA transcription algorithm is 99.48%. The 91 word type errors occur 347 times in the BAQ corpus which contains 77,430 words. Therefore, the computed accuracy of the automatic Arabic > IPA transcription algorithm is 99.55%.

On the other hand, contextual transcription of the BAQ corpus words is concerned with transcribing the words in context. They are transcribed so as to represent co-articulatory effects in continuous speech but with a definite pause at the end of each sentence. For example, the two sentences/verses in figure 3 " عَمَّ يَتَسَاءَلُونَ ﴿١﴾ عَنِ النَّبَإِ الْعَظِيمِ ﴿٢﴾" "About what are they asking one another? (1) About the great news - (2) " are transcribed contextually into /ʕamma jatasaːʔaluːn (1) ʕani nnabaʔi lʕaðˤijm (2)/. Verification of the contextual transcription tier of the BAQ corpus is done by checking these transcriptions sentence by sentence. The Quranic verses in the BAQ corpus are divided into sentences and pauses are defined as either major or minor. This information, which is already provided, makes our task simpler and more accurate.

78	1	1	1	عَمَّ	عَمَّ	P	-	ʕamma	ʕamma	Eam~a	
78	1	1	2	يَتَسَآءَلُونَ	يَتَسَاءَلُونَ	V	ǁ	jatasaːʔaluːna	jatasaːʔaluːn	yatasaA'aluwna	سأل
78	2	1	1	عَنِ	عَنِ	P	-	ʕani	ʕani	Eani	
78	2	1	2	ٱلنَّبَإِ	النَّبَإِ	N	-	ʔannabaʔi	nnabaʔi	Aln~aba<i	نبأ
78	2	1	3	ٱلْعَظِيمِ	الْعَظِيمِ	N	ǁ	ʔalʕaðˤijmi	lʕaðˤijm	AloEaZiymi	عظم
78	3	1	1	ٱلَّذِى	الَّذِي	N	-	ʔallaðiː	ʔallaðiː	Al~a*iy	
78	3	1	2	هُمْ	هُمْ	N	-	hum	Hum	humo	
78	3	1	3	فِيهِ	فِيهِ	P	-	fiːhi	fiːhi	fiyhi	
78	3	1	4	مُخْتَلِفُونَ	مُخْتَلِفُونَ	N	ǁ	muxtalifuːna	muxtalifuːn	muxotalifuwna	خلف
78	4	1	1	كَلَّا	كَلَّا	P	-	kallaː	kallaː	kal~aA	
78	4	1	2	سَيَعْلَمُونَ	سَيَعْلَمُونَ	V	ǁ	sajaʕlamuːna	sajaʕlamuːn	sayaEolamuwna	علم

Figure 4: shows a sample of the multi-tiered BAQ dataset version 2.0.

7. The multi-tiered *Boundary Annotated Qur'an* dataset: version 2.0

The *Boundary-Annotated Qur'an* dataset: version 1.0 contains 13 tiers, including: 2 tiers for the Arabic word, 2 tiers for part-of-speech, 2 tiers for boundary types, etc. Version 2.0 adds another 4 tiers for the BAQ dataset. These tiers are: (i) an IPA pausal transcription of the corpus words with case ending, (ii) an IPA contextual transcription tier, (iii) transliteration tier using Tim Buckwalter transliteration scheme[4], and (iv) root for each word in the dataset. Figure 4 shows a sample of the multi-tiered BAQ dataset version 2.0.

8. Conclusions

In this paper, we have extended the development of the Boundary Annotated Qur'an: a dataset for machine learning. Version 1.0 and 2.0 of the BAQ dataset contains multiple annotation tiers in a machine readable format. IPA phonetic transcriptions of the Qur'an are newly added tiers. Pausal phonemic transcriptions with and without case endings were automatically generated and added to the dataset. These transcriptions were then manually verified and corrected to reach 100% accurate dataset. A transliteration tier was added using Tim Buckwalter's transliteration scheme for MSA Arabic words. This shows the difference between 1-to-1 letter mapping and IPA phonetic transcriptions. Finally, the root of each word in the dataset was added.

9. References

Al-Ghalayyni. 2005. جامع الدروس العربية "Jami' Al-Duroos Al-Arabia" Saida - Lebanon: Al-Maktaba Al-Asriyiah "المكتبة العصرية".

Beckman, M. and Hirschberg, J. 1994. *The ToBI annotation conventions*. The Ohio State University and AT&T Bell Laboratories, unpublished manuscript. Online. Accessed September 2011. ftp://ftp.ling.ohio-state.edu/pub/phonetics/TOBI/ToBI/ToBI.6.html.

Brierley, C., Sawalha, M., Heselwood, B. and Atwell, E. (Under review). A Verified Arabic-IPA Mapping for Arabic Transcription Technology Informed by Quranic Recitation, Traditional Arabic Linguistics, and Modern Phonetics. Submitted to the *Journal of Semitic Studies*.

Brierley, C., Sawalha, M. and Atwell, E. 2014. Tools for Arabic Natural Language Processing: a case study in *qalqalah* prosody. To appear in *Proc. Language Resources and Evaluation Conference (LREC 2014),* Reykjavik.

Brierley, C., Sawalha, M., Atwell, E. 2012. "Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing." In *Proceedings of Language Resources and Evaluation Conference (LREC),* Istanbul, Turkey.

Habash, N., Soudi, A. and Buckwalter, T. 2007. On Arabic Transliteration. In Soudi, A., van den Bosch, A. and Neumann, G. (eds.), *Arabic Computational Morphology: Knowledge-based and Empirical Methods*.

Harrag, A. and Mohamadi, T. 2010. QSDAS: New Quranic Speech Database for Arabic Speaker Recognition. In *The Arabian Journal for Science and Engineering*. 35, 2C. 7-19.

Hassan, Z.M. and Heselwood, B. (eds.). 2011. *Instrumental Studies in Arabic Phonetics*. Amsterdam. John Benjamins Publishing Company.

Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge. Cambridge University Press.

Sawalha *et al* 2014a. (Forthcoming). IPA transcription technology for Classical and Modern Standard Arabic.

Sawalha, M., Brierley, C., and Atwell, E. 2012a. 'Predicting Phrase Breaks in Classical and Modern Standard Arabic Text.' In *Proceedings of LREC 2012,* Istanbul, Turkey.

Sawalha, M., Brierley, C., and Atwell, E. 2012b. "Open-Source Boundary-Annotated Qur'an Corpus for Arabic and Phrase Breaks Prediction in Classical and Modern Standard Arabic Text." In *Journal of Speech Sciences*, 2.2.

Sharaf, A.B. 2011. Automatic categorization of Qur'anic chapters. In *7th. International Computing Conference in Arabic (ICCA'11)*, Riyadh, KSA.

Taylor, L.J. and Knowles, G. 1988. 'Manual of Information to Accompany the SEC Corpus: The machine readable corpus of spoken English.‟ Accessed: January 2010.

Wells, J.C. 2002. *SAMPA for Arabic*. Online. Accessed: 25.04.2013. http://www.phon.ucl.ac.uk/home/sampa/arabic.htm

Wright, W. 1996. *A Grammar of the Arabic Language, Translated from the German of Caspari, and Edited with Numerous Additions and Corrections*. Beirut: Librairie du Liban.

[4] http://www.qamus.org/transliteration.htm

Appendix I: Arabic > IPA mapping

Arabic consonants	IPA symbol selection	Equivalent sound (if any) in English
ا	aː	bag
ب	b	big
ت	t	tin
ث	θ	thin
ج	ʤ	jam
ح	ħ	breathy ‘h’ as in hollow or whole
خ	x	loch
د	d	dog
ذ	ð	there
ر	r	rock
ز	z	zoo
س	s	sat
ش	ʃ	shall
ص	sˤ	a *bit* like the ‘so’ sound in ***so***ck
ض	dˤ	a *bit* like 'd' sound in '***d***uck', 'bu***d***', 'no***d***'
ط	tˤ	a *bit* like 't' sound in 'bough***t***', 'bo***tt***le'
ظ	ðˤ	no English equivalent but *voiced th*-like
ع	ʕ	no English equivalent
غ	ɣ	like the ‘r’ in the French word *rue*
ف	f	fun
ق	q	no English equivalent
ك	k	king
ل	l	lemon
م	m	man
ن	n	next
ه	h	house
و	w	will
ي	j	yellow
ء	ʔ	glottal stop as in Cockney *bottle*

Shaded cell: We are using /x/ for /χ/ for better readability of IPA transcriptions

Arabic short and long vowels	IPA	Equivalent sound (if any) in English
◌َ	a	short ‘a’ as in m*a*n
◌ِ	i	short ‘i’ as in h*i*m
◌ُ	u	short ‘u’ as in f*u*n
ا	aː	long ‘a’ as in c*a*r
ي	iː	long ‘i’ sound as in sh*ee*p
و	uː	long ‘u’ sound as in b*oo*t

Appendix II: A sample of the Qur'an frequency list with IPA transcriptions of pausal form with case ending

word type number	Word type frequency	Word type	Word type in IPA
1	1673	مِنْ	min
2	1185	فِي	fiː
3	1010	مَا	maː
4	828	اللَّهِ	ʔallaːhi
5	812	لَا	laː
6	810	الَّذِينَ	ʔallaðiːna
7	733	اللَّهُ	ʔallaːhu
8	691	مِنَ	mina
9	670	عَلَى	ʕalaː
10	662	إِلَّا	ʔillaː
11	658	وَلَا	walaː
12	646	وَمَا	wamaː
13	609	إِنَّ	ʔinna
14	592	اللَّهَ	ʔallaːha
15	519	أَنْ	ʔan
16	416	قَالَ	qaːla
17	405	إِلَى	ʔilaː
18	372	مَنْ	man
19	344	إِنْ	ʔin
20	337	ثُمَّ	θumma
21	327	بِهِ	bihi
22	325	لَهُمْ	lahum
23	323	كَانَ	kaːna
24	296	بِمَا	bimaː
25	294	لَكُمْ	lakum
26	280	ذَلِكَ	ðaːlika
27	275	لَهُ	lahu
28	268	الَّذِي	ʔallaðiː
29	265	هُوَ	huwa
30	264	أَوْ	ʔaw
31	263	قُلْ	qul
32	253	آمَنُوا	ʔaːmanuː
33	250	قَالُوا	qaːluː
34	241	فِيهَا	fiːhaː
35	239	وَاللَّهُ	wallaːhu
36	234	وَمَنْ	waman
37	229	كَانُوا	kaːnuː
38	219	الْأَرْضِ	ʔalʔardˤi
39	195	إِذَا	ʔiðaː
40	190	هَذَا	haːðaː