Kamal Abou Mikhael
LREC 2014 MEDAR Grant Report
September 10, 2014

Introduction

LREC 2014 was a great experience that exposed me to interesting research activity in mostly unfamiliar areas. What follows is a report on the sessions that I attended, consisting of poster presentations and oral presentations. While at LREC 2014 I attended one pre-conference workshop, the main conference, and a post-conference workshop in which I presented a paper. I am very grateful to MEDAR for having afforded me this opportunity to travel to Iceland for LREC 2014, it was a truly unique and eye opening experience that allowed me to encounter new ideas and knowledgeable people. This trip and the experience as a publisher and presenter was highly beneficial for the project I presented, `http://greekarabicnt.org`. Prior to the conference, the prospect of publishing accelerated my progress; at the conference, I met with many interested researchers who gave me great ideas; after the conference, I met more researchers who were absent from the conference to whom I was introduced by those I met at the conference.

Pre-Conference Workshop - Tuesday, May 27

Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT)

King Abdullah Initiative for Arabic Content – Prof. Mansour Algamdi (Key note Speech)
The key note speaker was Dr. Mansour Algamdi who presented the King Abdullah Initiative for Arabic Content. The vision of the initiative is to enrich the Arabic content of the web and provide tools to benefit from that content. The initiative operates on a three-level model consisting of (1) hardware and human resources, (2) software and literacy/culture which serves language, and (3) content. The goal is to make available databases that can be used by systems that perform generation, synthesis, recognition, analysis, scoring, translation, classification, search, encryption, and compression. Among such databases made available through the initiative are KACST Arabic Phonetic Database (KAPD), Quran Database, KACST Diacritized Text, Arabic Corpus, Standard Arabic Single Speaker Corpus, and KACST Parallel Translated Text corpus of Arabic and Hebrew.

The talk stirred up inquiries about the availability of such sources. The sources are in the process of being made available through Sourceforge. The Arabic dictionary and morphologizer are available at Sourceforge. The KACST Diacritized Text is available through contact with Dr. Algamdi. The Arabic Corpus is not available fully, a small part of it (220,000 tokens) is available along with information about the entire copurs such as frequency lists. This initiative is of high value and the use of the resources, once made fully available, is bound to move forward Arabic NLP.

Critical Survey of the Freely Available Arabic Corpora – Wajdi Zaghouani
Wajdi Zaghouani (Carnegie Mellon University Qatar) presented a paper titled Critical Survey of the Freely Available Arabic Corpora. Similar to the key note speech, this work addressed the necessity of linguistic data for building systems. However, it also stressed that free data is also important to researchers whose institutions do not have paid memberships to institutions that provide Arabic language data. This initiative consisted of a survey that was sent out to various institutions, eliciting from them information about free Arabic language resources they have. A first round of surveys involved 26 participants and resulted in 66 resources being identified. The current set of identified free corpora cover the following major categories of corpora: raw text corpora, annotated corpora, lexicon, speech corpora, handwriting recognition corpora, and miscellaneous corpora types (questions/answers, comparable corpora, plagiarism detection, and summaries). However, there is room for improvement for data availability in the areas of speech corpora, handwriting corpora, and dialectical corpora. The online survey can be found

at http://tinyurl.com/l63fve8 and the current list can be found at https://www.qatar.cmu.edu/ wajdiz/corpora.html.

Automatic Readability Prediction for Modern Standard Arabic – Jonathan Forsyth

Jonathan Forsyth (Brigham Young University) presented a paper on Automatic Readability Prediction for Modern Standard Arabic. It observed that although machine learning based methods have been applied to readability prediction in English, they have yet to be applied to Arabic. It introduced such a method for automated prediction of MSA readability. The method was applied to a 179 document corpus (67,532 tokens) in the following manner: (1) the corpus was annotated for Interagency Language Roundtable (ILR) level, (2) lexical and discourse features were extracted with a Perl script (162 features were used in the experiments), and (3) the Tilburg Memory-Based Learning (TiMBL) program was used to predict the ILR level of the document based on the features. The prediction was tuned with 10-fold cross validation. 3-way classification achieved an average F-score of 0.719 and 5-way classification achieved an average F-score of 0.519. Future improvements proposed include a larger corpus for training the classifiers and using of Arabic syntactic features. An enhanced system would be the basis of a pedagogical tool that would allow students and instructors to know the readability of a given document.

Subjectivity and Sentiment Analysis of Arabic Twitter Feeds with Limited Resources – Eshrag Refaee and Verena Rieser

Eshrag Refaee and Verena Rieser (Interaction Lab and Heriot-Watt University, respectively) presented Subjectivity and Sentiment Analysis of Arabic Twitter Feeds with Limited Resources. This work carried out automatic Subjectivity and Sentiment Analysis (SSA) on tweets written in various Arabic dialects. The researchers compiled their own set of training data consisting of 3K of twitter data, manually annotated for SSA features (positive, negative, or neutral), morphological features, syntactic features, and semantic features. The machine learning model that achieved the best results was the Support Vector Model (SVM). Various experiments were run combining various types of features and various means of SSA evaluations (positive vs. negative, polar vs. neutral, and positive vs. negative vs. neutral). Polar vs. neutral yielded the most consistent F-scores and percentage accuracies. This first validation employed 10-fold cross-validation and achieved accuracy scores of around 87%. A second evaluation that was carried out on an independent data set, but only achieved an accuracy of around 39%. An error analysis revealed that it was due to a topic shift. Thus, it is hypothesized that better results can be achieved when a larger set of data with a wider set of topics is used in the training phase.

Large Arabic Web Corpora of High Quality: The Dimensions Time and Origin – Thomas Eckart, Uwe Quasthoff, Faisal Alshargi and Dirk Goldhahn

Thomas Eckart, Uwe Quasthoff, Faisal Alshargi and Dirk Goldhahn (University of Leipzig) presented Large Arabic Web Corpora of High Quality: The Dimensions Time and Origin. This work tackles the problem of the variety among Arabic dialects and seeks to employ a "broad and steady text acquisition approach". This approach is used to compose the Arabic part of the Leipzig Corpora Collection (LCC). The work used various web crawlers and web extraction techniques and focused on special domains. The extracted text was processed to form a corpus. Sentence scrambling was used to prevent violation of, at the least, German copyright laws. This corpus includes country specific newspaper based corpora for 17 Arab speaking countries. The data allows for comparisons among variants within and between Arabic dialects as well as diachronic analysis. The corpus is continuously growing and its data is available through various web-based interfaces. Some issues observed throughout the discussion in the workshop included the handling of the "al-" definite article, which varied depending on the application and the definition of the sentence as determined using a rule-based approach.

An Algerian Arabic / French Code-Switched Corpus – Ryan Cotterell and Chris Callison-Burch

Ryan Cotterell, Adithya Renduchintala, Naomi Saphra, and Chris Callison-Burch (Johns Hopkins University and University of Pennsylvania) presented An Algerian Arabic / French Code-Switched Corpus.

This work addressed the problems of certain properties missing among current Arabic corpora, namely the lack of non-MSA, romanized, code-switched corpora. The data sought in this work also presented the problem of analysis brought on by romanized Arabic and the problematic nature of code switching. Comments from the Al-Sharq newspaper were extracted, resulting in over 6.5 million tokens. The data was observed to contain variations in spelling and grammar, elongations, abbreviations, and features of web text communication such as emoticons and chat-specific abbreviations. The comments were found to be 68% French, 30% Arabic, and 2% other. One of the problems faced during annotation was making the distinction between code switching and words borrowed from other languages. Among the future possibilities for this work is the creation of corpora for other bilingual Arabic countries as well as non-Arabic speaking countries such as Canada. Such corpora can be subject to diachronic and geographical analysis.

The workshop ended with a panel discussion regarding the current drawbacks of the available Arabic corpora. There was a brief mention of the desire for more data and more dialects to be represented. However, the main issue was regarding the difficulty of processing dialectical data. It was suggested that perhaps dialects can be standardized. Such efforts, though, may prove unsuccessful in light of the political implications of this issue and people's willingness to accept a standard. Nevertheless, the introduction of the idea and of samples conforming to a set of given standards may serve to help guide transcriptions of dialectical data.

Main Conference

Day 1 - Wednesday, May 27, 2014

Opening Ceremony

The opening session consisted of presentations by UNESCO and Icelandic officials (including UNESCO Director General Her Excellency Madame Irina Bokova, former President of the Iceland, UNESCO Goodwill Ambassador for Languages Madame Vigdis Finnbogadottir, and Mayor of Reykjavik Jon Gnarr) and ELRA officials (Presiden Nicoletta Calzolari and Secretary General Khalid Choukri). The presentations focused on the importance of promoting multilingualism and preserving local languages in a globalized, connected, and information driven world. The inclusion of languages that are on the periphery was seen as a means of leveling the playing field. It was stressed that language resources and research would play a crucial role in fulfilling this mission.

Introductory Session

The introductory session was chaired by ELRA Honorary President Joseph Mariani and consisted of the following key presentations:

A World of Language Resources by Marta Nagy-Rothengass

LRs are important as a social asset, economic asset, and opportunity. Sustainable LRs are valuable to the digital transformation, cross-lingual and cross-border collaboration, and the big data paradigm of research. LRs also have multiple dimensions: legal, application, social, technology, and economic. The potential and key role of LRs coupled with their multi-dimensionality requires a functional to allow them to be effective in diminishing language barriers.

Language Resources in the Data Economy - Nicoletta Calzolari and Khalid Choukri

There are various traditional activities centered around the roles and interaction btween between LR providers and users. There are also various types of LRs. What is desired is a basic language kit for all; its components must be identified and catalogued appropriately. When it comes to digital LRs, there are legal issues and issues of accessibility that in turn have to be articulated to policy makers. Interaction

with policy makers can be complicated especially for creating copyright regulations that are compliant with digital formats.

From the beginning, ELRA has realized the importance of LRs and the issues surrounding them as is evident in the history of LREC. ELRA is already active in helping with the sustainability and development of the state of LRs. To help LR providers, ELRA is aiming to create a licensing wizard that helps generate the appropriate license in light of various possibilities and implications for different types of LRs. The Language Resources and Evaluation Journal, co-edited by Nicoletta Calzolari, is a first of its kind and is bound to aid in reporting and enhancing the state of LR practices. ELRA is also participating in key initiatives such as the Thomson Reuters Conference Proceeding Index and the LRE Map. ELRA is also promoting recreatability, a translation corpus initiative, a universal LR identifications scheme, and more resource sharing. ELRA's efforts also serve to support open science by dealing with issues of reproducibility and attribution/citation in addition to efforts to create an open scientific information space and to involve funding agencies.

LREC 15th Anniversary Celebration - Joseph Marian The celebration of the 15th anniversary of LREC marks an era of many activities, milestones, and conferences that have featured significant contributions by many participants in papers, presentations, and poster sessions. There are many ambitions and possibilities for the future not only in terms of research, but also for installing permanent LREC community.

Poster Session P1 – Crowdsourcing (11:35-13:15)

Can the Crowd be Controlled?: A Case Study on Crowd Sourcing and Automatic Validation of Completed Tasks based on User Modeling – Balamurali A.R

Balamurali proposes a means of validating the quality of work in crowd sourcing. He observes that workers who return to the same reviewer for the same task can be automatically considered high quality. A framework based on this notion was developed and applied on an SMS annotation task. The results validated the hypothesis for the most part; there was a case where this approach attributed greater accuracy to a worker than his/her performance deserved.

When Transliteration Met Crowdsourcing : An Empirical Study of Transliteration via Crowdsourcing using Efficient, Non-redundant and Fair Quality Control – Mitesh M. Khapra, Ananthakrishnan Ramanathan, Anoop Kunchukuttan, Karthik Visweswariah and Pushpak Bhattacharyya

Mitesh et. al. explore quality control mechanisms for transliteration done via crowdsourcing. They attempt to avoid methods that are unfair to workers and expensive to requesters. They propose an automated rule-based evaluation of transliteration that is based on the mapping of consonants between the source language and the transliterated language. This approach was tested on Hindi to English transliteration and it was found to outperform QC methods such as consensus and sampling.

Collaboration in the Production of a Massively Multilingual Lexicon – Martin Benjamin

Benjamin presents the lexicon building approaches used by the Kamusi project, a initiative to build a multilingual lexicon. Building it draws from experts, existing data, ordinary speakers, and crowdsourcing. Techniques used include data mining, rewarding helpful input according to its worth, creating useful tools for users to provide input, and collecting user input through games. Crowd sourcing is used for translations, definitions, data merging, and validation.

Can Crowdsourcing be used for Effective Annotation of Arabic? – Wajdi Zaghouani and Kais Dukes

Zaghouani and Dukes tested the feasibility of crowdsouring for Arabic annotation. They propose an initial screening test to measure the competence and qualification of annotators. The Quranic Arabic Corpus was used as a gold standard. The tasks requested in the study were POS tagging and case ending identification. POS tagging was found to be easier than identifying case endings. Crowdsourcing was

found to be less effective than other methods for accomplishing these tasks.

Session P11 – MultiWord Expressions and Terms (14:45-16:25)

Linked Open Data and Web Corpus Data for noun compound bracketing – Pierre Andre Menard and Caroline Barriere
Menard and Barriere propose the use of Linked Open Data resources (DBPedia) for the task of noun compound bracketing as opposed to the conventional use of Web Corpus Data (Google Web/Books Ngrams). DBPedia is found to perform slightly below, but comparably to the Google Ngrams data. The authors propose a hybrid model.
Identifying Idioms in Chinese Translations – Wan Yu Ho, Christine Kng, Shan Wang and Francis Bond
Ho et. al. observe that English-to-Chinese translation often results in idiomatic expressions in the target language where there is none in the source. In order to further explore this, they built a lexicon of Chinese idioms (Chengyu) and explored their occurrences over four genres (Story, News, Essay, and Tourism), each represented by a corpus for the purpose of predicting translation trends. Each of the genres were shown to have distinct distribution profiles.
Collocation or Free Combination? – Applying Machine Translation Techniques to identify collocations in Japanese – Lis Pereira, Elga Strafella and Yuji Matsumoto
Pereira et. al. experiment with the notion that collocation can be distinguished from free combinations using machine translation systems. Since free combinations are usually translated word-for-word and collocations are not, MT is suspected to distinguish between the two. This approach was tested using English-Japanese corpora and dictionaries; it resulted in improved precision when run on a small set of data.
Building a Crisis Management Term Resource for Social Media: The Case of Floods and Protests – Irina Temnikova, Andrea Varga and Dogan Biyikli
Temnikova et. al. investigate the unique terminology used on Twitter to describe crises. The usefulness of this exploration lies in the ability to provide Crisis Management agencies with information about such events. Two types of crises were investigated, floods and protests, and four instances for each. Manual annotation showed that Twitter data differed from traditional data regarding such events with respect to both terminology and POS data. Various methods for automated identification of emergency events were tested (TermRaider, TF-IDF, and C-Value); all of them performed poorly. However, supervised learning was able to identify tweets inviting for action (action tweets) with 80% accuracy.

Session P16 – Morphology (16:45-18:05)

MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic – Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow and Ryan Roth
Pasha et. al. present a system for Arabic morphological analysis and disambiguation named MADAMIRA. It combines two previous works, (1) MADA, a morphological analyzer and disambiguration, and (2) AMIRA, which contains a tokenizer, POS tagger, and a shallow syntactic parser. The architectural design of MADAMIRA is based on that of MADA and contains components derived from AMIRA at the end of the processing (NER and syntactic analysis). MADAMIRA enjoys all the benefits of a Java implementation and is more than an order of a magnitude faster than its predecessor.

Session P17 – WordNet (16:45-18:05)

Mapping WordNet Domains, WordNet Topics and Wikipedia Categories to Generate Multilingual Domain Specific Resources – Spandana Gella, Carlo Strapparava and Vivi Nastase

Gella et. al. produce domain-specific and multilingual corpora by mapping WordNet topics to WordNet domains which are mapped to Wikipedia category hierarchies. This resulted in domain specific corpora for English, French, German, Italian, Portuguese, and Spanish. These corpora were further used to build a domain-specific sentiment lexicon.

Dense Components in the Structure of WordNet – Ahti Lohk, Kaarel Allik, Heili Orav and Leo Vohandu

Lohk et. al. observe that WordNets are highly useful for NLP applications and pay attention to their polysmey feature. With respect to that, they observe that polysemy can either complicate or enhance NLP applications based on its quality. They in turn redefine regular polysemy (RP) as a minimum of two sysnets having a minimum of two hypernyms with similar relations between them. To evaluate RP, they propose a dense component graph as a test pattern to help lexicographers evaluate inconsistencies in polysemy. They implemented this on versions 66 and 67 of the Estonian WordNet (EstWN) as well as the Princeton WordNet 3.1; 88% of the dense components identified required correction.

The Making of Ancient Greek WordNet – Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini and Gregory Crane

Bizzoni et. al. present a paper about the AncientGreekWord-Net, namely how it is built and checked. It is built using Greek-English dictionaries from which definitions are extracted with Greek synonyms being determined by previously established synonymity among their English translations. Validation consisted of filtering out polysemy established by colloquial English terms followed by manual correction via a graphical user interface. The data will be made available as (Linguistic) Linked Open Data.

Session P18 – Corpora and Annotation (18:10-19:30)

A Multidialectal Parallel Corpus of Arabic – Houda Bouamor, Nizar Habash and Kemal Oflazer

Bouamor et. al. report on their construction of parallel corpus consisting of Standard Arabic, English, and the following qolloquial dialects: Egyptian, Tunisian, Jordanian, Palestinian, and Syrian. This is a previously unavailable resource consisting of 2,000 sentences for each variety. Each Arabic dialect is represented with near or over 11,000 tokens and averages near or above 9 tokens per sentence. The data confirmed expectations regarding degree of similarity between the represented varieties. The work will be extended with additional information such as morphological analysis, POS tags, and manual word-by-word alignment.

The American Local News Corpus – Ann Irvine, Joshua Langfus and Chris Callison-Burch

Irvine et. al. report on the American Local News Corpus (ALNC) which contains 4 billion words and represents all 50 U.S. states and nearly 2,000 cities. The data is tagged with time and location data. This corpus facilitates analysis across the dimensions of location and time; this was showcased with two analyses regarding sports and capital murder. Further proposed uses are for fields such as public health, social science, sociology, and journalism.

Session P19 – Document Classification, Text Categorisation (18:10-19:30)

Detecting Document Structure in a Very Large Corpus of UK Financial Reports – Mahmoud El-Haj, Paul Rayson, Steve Young and Martin Walker

El-Haj et. al. tackle the problem of detecting document structure in UK financial reports for the purpose of determining what affects quality of reporting. The analysis is at the section level across various documents rather than within a single document. Analysis was conducted across the dimensions of time, companies, and sectors. Analysis focused on metrics, such as readability, hedging, and forward-looking language.

Sockpuppet Detection in Wikipedia: A Corpus of Real-World Deceptive Writing for Linking Identities – Thamar Solorio, Ragib Hasan and Mainul Mizan

Solorio et. al. take on the problem of sockpuppet detection. Sockpuppets are fake user accounts used for the purpose of publishing false data and carrying out abusive communication. The corpus consists of Wikipedia content that has been called into question. The authors used a machine learning approach depends on features such as punctuation, character frequency, word frequency, timing, and word distances. F-measures of 73% were reached when using all features.

Day 2 - Thursday, May 28, 2014

Keynote Speech: Language Technology for Commerce, the eBay Way – Hassan Sawwaf

The second day opened with a keynote speech by Hassan Sawwaf of eBay spoke about eBay's global mission and his work with in-house machine translation systems to serve that mission. Globalized eBay requires translation due to geographic expansion; translation is used for the purposes of crosslingual IR, user-to-user communication, and displaying titles and descriptions. Third party tools had served eBay well for descriptions and well-formed queries, but it faced many challenges when it came to ungrammatical input and named entities. However, in-house technology has resulted in improvements. Challenges include the employment of mathematical models and end-tail sparseness. Enhancing eBay MT requires learning from users through explicit feedback (ratings), implicit feedback (purchases), and other behavioral data. It also depends on human evaluation from in-house linguists and outside evaluators; it also employs 10,000 to 20,0000 translations, having 3-5 evaluators per sentence. The Q&A session consisted of two questions and revealed that eBay was not focusing on speech input and that the challenge of aligning categories (across languages) in eBay translation (and in general) requires help from the field of machine learning.

Session O20 – Grammar, Lexicon and Morphology (9:45-11:25)

The Interplay Between Lexical and Syntactic Resources in Incremental Parsebanking – Victoria Rosen, Petter Haugereid, Martha Thunes, Gyri S. Losnegaard and Helge Dyvik
Rosen et. al. addresses the problem faced when parsing depends on lexicons that has missing lexical or morphological information for certain words. Incremental parsebanking is used to insert missing information either before or after parsing which results in a richer lexicon and fully informed parsing.
An Efficient Language Independent Toolkit for Complete Morphological Disambiguation – Laszlo Laki and Gyorgy Orosz
Laki et. al. propose a language independent system for morphological disambiguation. This statistical machine translation (SMT) system was compared with others that are based on hidden markov models (HMM) for languages such as Hungarian, Croatian, and Serbian. The SMT toolkit outperformed all HMM systems for the task of POS tagging, but outperformed only one of them in the task of lemmatization.
Language Resource Addition: Dictionary or Corpus? – Shinsuke Mori and Graham Neubig
Mori & Neubig compare two methods of addressing word-segmentation and POS tagging for language resources in Japanese. The approach of adding dictionary entries was compared to adding annotated sentences to a training corpus. Addition of annotated sentences was found to be more effective.
Utilizing Constituent Structure for Compound Analysis – Kristin Bjarnadottir and Jon Dadason
Bjarnadottir and Dadason presents a toolkit to identify unknown compounds in Icelandic, a language in which compounds are very common. The method stands in contrast to parallel corpus methods, frequency-based methods, and ensemble methods. Possible word segmentations were represented as binary trees and probabilities were estimated bottom-up; the tree with the highest probability was chosen as the solution. The method allowed the choosing of various levels of granularity.

Word Semantic Similarity for Morphologically Rich Languages – Kalliopi Zervanou, Elias Iosif and Alexandros Potamianos
Zervanou et. al. employ selective stemming based on a computed measure of semantic distortion in order to improve estimates of semantic similarity. This methods selects rules for stemming in light of semantic distortion. The method was attempted on Greek and German, chosen for their rich morphology, and English, chosen for its simpler morphology. For each language, the algorithm was run using co-occurrence and context-based metrics. Co-occurrence metrics yielded better results for English whereas context-based metrics yielded better results for Greek and German.

Session O24 – Document Classification (11:45-13:05)

Cross-Language Authorship Attribution – Dasha Bogdanova and Angeliki Lazaridou
Bogdanova and Lazaridou presented two solutions for cross-language authorship, one based on machine translation (MT) and the other based on high level features (mostly cross-lingual). The high level features (HLF) were related to sentiment, emotions, perception, and average sentence length; these were cross-lingual. POS frequencies were also considered among HLFs, but not all features in this category were cross-lingual. The HLF based method did not perform comparably to the MT method; however, a combination of MT and HLF performed comparably to MT methods. The authors posited that better LRs were needed to improve the performance of HLF methods.
Learning from Domain Complexity – Robert Remus and Dominique Ziegelmayer
Remus and Ziegelmayer explored the relation between the performance of sentiment analysis methods and the complexity of the domain of the text. Domain complexity was measured with percentage of rare words, word richness, relative entropy, and corpus homogeneity. Their experiments showed that performance of ML based sentiment analysis methods could be predicted by domain complexity as calculated with the aforementioned measures. Domain complexity was also found to be useful in choosing between conservative and aggressive n-gram feature selection.
Using Word Familiarities and Word Associations to Measure Corpus Representativeness – Reinhard Rapp
Rapp proposes an approach to corpus reprasentativeness that compares statistics about word associations and frequencies to statistics based on human judgments about word association norms and word familiarity norms. Higher similarity between the two sets of statistics implies greater reprasentativeness. This approach is applied to the following corpora: Brown, VNC, UkWaC, Wikipedia, and Gigaword; BNC was found to have the greatest reprasentativeness among these. One of the identified shortcomings of the methods was that it did not consider higher level features. Future plans include extending the methodology to second order associations via WordNet.
A Modular System for Rule-based Text Categorisation – Marco Del Tredici and Malvina Nissim
Tredici and Nissim explore the effectiveness of an enhanced rule-based approach to text categorisation that employs a hierarchical structure of rules that allows rule reusability. Small sets of rules are used to create atomic categories and aggregate rules. The system was evaluated with a use case on Islamic terrorism and rule reusability was evaluated with a use case on Christian terrorism. The system's performance was found to be comparable to that of currently used machine learning models. Reusability was achieved efficiently due to the minimal rule rewrites.

Invited Talk: Icelandic Quirks: Testing Linguistic Theories and Language Technology – Porhallur Eyporsson

In the middle of the day, this special session with a local flavor was held. It showcased the many unique features of the Icelandic language that made is stand out from other languages and, more importantly,

contradict the expectations and generalities derived by theoretical linguists. Oblique subjects showed that dative pronouns can be the subject, showing that there was no one-to-one relation between case and grammatical relations. Stylistic fronting broke down the distinction between heads and phrases (e.g., "elections, which place have taken"). The new passive showed that the accusative can be preserved with a passive verb. The talk also put some perspective on the state and uniqueness of Icelandic. The language was said to be characterized by both archaisms and innovations. Innovations were found in phonology, but less frequently in syntax. Corpus linguistics showed that modern Icelandic occurs with low frequency. Finally, oblique subjects, although discovered first in Icelandic, were also found in Indo-European. What could be said conclusively, though, is that Icelandic is the language in which the quirks were first discovered conclusively.

Session P34 – Corpora and Annotation (14:55-16:35)

Large Scale Arabic Error Annotation: Guidelines and Framework – Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani and Kemal Oflazer

Zaghaouani et. al. introduce (1) a web-based annotation framework, (2) annotation guidelines, and (3) annotator training issues from their experience in constructing an Arabic corpus of errors and corrections as part of the Qatar Arabic Language Bank (QALB). Such a corpus can be used for training models and as a gold standard. The errors were related to spelling, punctuation, word choice, morphology, syntax, proper names, and dialect usage. Annotation guidelines were revised after an initial training phase; revised guidelines were used in the final production phase. Annotation was evaluated in terms of quality of input and inter-annotator agreement. MADA, the morphological analysis and disambiguation tool, was able to improve input quality by 10%. Inter-annotator agreement was rather high as a result of the guidelines.

Session P36 - Metaphors (14:55-16:35)

A Multi-Cultural Repository of Automatically Discovered Linguistic and Conceptual Metaphors – Samira Shaikh, Tomek Strzalkowski, Ting Liu, George Aaron Broadwell, Boris Yamrom, Sarah Taylor, Laurie Feldman, Kit Cho, Umit Boz, Ignacio Cases, Yuliya Peshkova and Ching-Sheng Lin

Shaikh et. al. describe an on going project to build a multi-lingual repository of metaphors that so far includes American English, Mexican Spanish, Russian, and Iranian Farsi. The metaphors were extracted from unrestricted text. A subset of the data was validated using social science experiments. Among the benefits of the repository are (1) a metaphor based cross-cultural view of issues of governance and economic inequality and (2) recognition of metaphor patterns.

Two Approaches to Metaphor Detection – Brian MacWhinney and Davida Fromm

MacWhinney and Fromm compare two common metaphor detection techniques for the target domain of economic inequality. The first is based on violation of selectional preferences or selectional restrictions methods; this was evaluated with various implementations (based on semantic features, WordNet, VebNet, and an ontological database). The first suffered because it depended on the quality of the available linguistic tools and resources. The second is corpus-based analysis where grammatical relations detected using SktechEngine allowed for WordScketches to be detected for words related to the target domain. The corpus-based WordSketch implementation far outperformed the selectional restriction methods with a performance that was triply as good in most instances with respect to both recall and precision.

Mining Online Discussion Forums for Metaphors – Andrew Gargett and John Barnden

Gargett and Barnden combine both manual and automatic methods to detect metaphors in the domains of illness and conflict. They use the Metaphor Identification Procedure (MIP(VU)) to manually annotate

the data. The manual annotation is supported by a binary confidence measure (0 or 1) that takes into consideration an input domain, an output domain, and a set of features determining confidence. Among the future aims of the authors are automation of metaphor detection as well as the generation of metaphor.

Session O32 – Parallel Corpora (16:55-18:15)

Creating a Massively Parallel Bible Corpus – Thomas Mayer and Michael Cysouw
Mayer and Cysouw discuss their efforts in building a massive parallel corpus using translations of the Bible, a book that has, is, and will be translated into new languages. The corpus at the time of writing included over 900 translations in over 830 languages. Text was collected from various publicly available sources such as online repositories and websites. However, the data is only partially available due to copyright restrictions on the texts. The data has been processed (tokenization and Unicode normalization). The project makes available co-occurence information (as sparse matrices) as well as word frequency lists.
DCEP -Digital Corpus of the European Parliament – Najeh Hajlaoui, David Kolovratnik, Jaakko Vayrynen, Ralf Steinberger and Daniel Varga
Hajlaoui et. al. present the Digital Corpus of European Parliament (DCEP) which contains documents of 23 languages (1.37 billion total words) and a diverse set of subject domains and document types from over 10 years (2001-2012). They discuss issues of acquisition and processing. Sentence-level alignment is not complete, however, what has been accomplished has been shown to be useful in statistical machine translation.
Innovations in Parallel Corpus Search Tools – Martin Volk, Johannes Graen and Elena Callegaro
Volk et. al. give an overview of search systems used for parallel corpora and argue that automated word alignments allow better search results than those based on sentence based corpora. Furthermore, the authors propose the development of tools that conduct search based on parallel tree banks.
An Open-Source Heavily Multilingual Translation Graph Extracted from Wiktionaries and Parallel Corpora – Valerie Hanoka and Benoit Sagot
Hanoka and Sagot introduce an open-source multilingual translation database consisting of 664 languages named YaMTG (Yet another Multilingual Translation Graph). The resource draws its data from multiple wiktionaries as well as OPUS parallel corpora. Data was extracted and poor translations were eliminated with generic and graph-based heuristics; the resulting quality for all data was 91% correctness and reached 97.2% for wiktionaries. Additional translations and tweaked thresholds in the graph-based filtering are expected to further improve the achieved results.

Session P47 – Language Identification (18:20-19:20)

Vocabulary-Based Language Similarity using Web Corpora – Dirk Goldhahn and Uwe Quasthoff
Goldhahn and Quasthoff approach automated language similarity based on similarity of vocabulary. They used word-based approaches as well as approaches based on parallel text. In word based approaches, a profile was built for each language (based on frequency of word or letter trigrams) then the similarity of the profile was compared, resulting in a similarity matrix from which language clusters were created. In this approach, tri-grams were found to be a more reliable basis for a language profile than words. Parallel text approaches were based on cross-language word distributions across sentences. Word pairs were identified and cognates were also identified both on the phonetic and orthographic levels; phonetic cognates were found to be superior. This approach was found to capture a similarity between languages that is based on grammar. Phonetic cognates were found to be the best basis for identifying similar languages.

Automatic Language Identity Tagging on Word and Sentence-Level in Multilingual Text Sources: a Case-Study on Luxembourgish – Thomas Lavergne, Gilles Adda, Martine Adda-Decker and Lori Lamel
Lavergne et. al. take on the problem of the scarcity of language resrouces for Luxembourgish, a language used with much code switching. They propose (1) a design for a manually annotated language of code-switched sentences and (2) tools for extracting such sentences. The tools employ Maxent and linear chain CRF. The tools were applied to web data to build lexicons and language models which were used for Automated Speech Recognition system for Luxembourgish. The system achieved 25% WER on the Quaero development data. The resulting freely available corpus consisted of 924 sentences.
VarClass: An Open-source Language Identification Tool for Language Varieties – Marcos Zampieri and Binyam Gebre
Zampieri and Gebre developed VarClass, an open-source tool for language identification (with GUI). In contrast to other similar tools, it focuses on language varieties. So far, it models 27 languages, 10 of which are language varieties. The tool was reported to reach 90.5% accuracy. The most highly featured among the models is Spanish (4 varieties) followed by English, French, and Portuguese (2). Best results among these languages were achieved for French followed, in order, by Portuguese, Spanish, and English.
On the Romance Languages Mutual Intelligibility – Liviu Dinu and Alina Maria Ciobanu
Dinu and Ciobanu use lexical similarity to cluster languages. At the core of the approach is cognate identification that is based on etymology detection (extracted from dictionaries). The approach was tested on the romance languages (Romanian, Italian, French, Spanish and Portuguese) using three parallel corpora: "1984" by George Orwell, Europarl, and Wikipedia content. "1984" and Europarl were found to be close to expected results; Wikipedia content was proposed to have strayed from expectations due to Wikipedia not being published as a parallel corpus. The authors intend to extend the work to apply it to the similarities of language families.

Day 3 - Friday, May 29, 2014

Keynote Speech

The third day opened with a keynote speech ("When will robots speak like you and me?") by Luc Steels. It presented a new method that teaches robots to speak through an incremental interactive process that employs a tutor/learner model which is an alternative to corpus and statistical approaches. Currently, this process consists of a naming game invention/adoption strategy whereby the speaker generates a phonetic/phonmeic label for objects and actions and a hearer aligns its understanding with that of the speaker. The naming and alignment are based on visual experience with semiotic network linking sensory experience to prototypes (individuals or words). The keynote speech was made highly interactive with both private and public video footage of the cutting-edge research activity. The project has the ambition to scale up from naming to grammar.

Session O39 – Information Extraction (2) (9:45-10:25)

Combining Dependency Information and Generalization in a Pattern-based Approach to the Classification of Lexical-Semantic Relation Instances – Silvia Necsulescu, Sara Mendes and Nuria Bel
Necsulescu et. al. addresses the problem of classifying instances of lexical semantic relations. They explore the relations of hypernymy, co-hyponymy, meronymy, attributes of nouns, and actions done by or to a noun. They form a Pattern-based Classification Model which employs two approaches: (1) information from dependency paths of up to three edges and (2) generalization using part of speech information. The combination of the two approaches yielded the best results when compared to other systems as indicated by the F-measure scores.

Corpus and Method for Identifying Citations in Non-Academic Text – Yifan He and Adam Meyers
He and Meyers presented a solution for identifying citations in non-academic text. In contrast to the systematicity and regularity of academic text, non-academic text contain ad-hoc citations. The type of non-academic text investigated was patents. They employed a CRF classifier that was trained using manually annotated documents. They were able to achieve an improvement of 0.03 in the F-score (from 0.8 to 0.83) by re-ranking citations based on non-local information (matching braces and ration of alpha-numeric characters).

Session O38 – Paraphases (10:45-11:25)

Creating and Using Large Monolingual Parallel Corpora for Sentential Paraphrase Generation – Sander Wubben, Antal van den Bosch and Emiel Krahmer
Wubben et. al. propose machine translation as a means of generating paraphrases. They used a parallel corpus of English and Dutch constructed from aligned headlines to train a PBMT-based model that employs re-ranking. They compared the results against a word-substitution baseline and showed that the system achieved superior results based on (1) human judgments and (2) the NIST and BLEU MT metrics. They observed that, when compared to the baseline system, the performance of the PBMT-R system degraded less as it deviated from the original wording.
Aligning Predicate-Argument Structures for Paraphrase Fragment Extraction – Michaela Regneri, Rui Wang and Manfred Pinkal
Regneri et. al. consider the use of fragments as the unit of paraphrase extraction rather than entire sentences, considering them too specific. Their approach prunes sentence paraphrases to phrase-level units. The paraphrase fragments are extracted based on semantic roles, which results in the selection of paraphrases that vary semantically and syntactically when compared to other systems. The performance of the system was on-par with two other systems (Giza-Baseline and VP-baseline); it also extracts a smaller number of unrelated fragment pairs.

Session O44 – Grammar and Parsing (2) (11:45-13:25)

When POS Data Sets Don't Add Up: Combatting Sample Bias – Dirk Hovy, Barbara Plank and Anders Sogaard
Hovy et. al. address the issue of lack of homogeneity in annotation among Twitter corpora for the purpose of combining such corpora. The dissimilarities are with respect to tag sets, data in the tag sets once unified, data bias, and label bias. They conducted experiments using heterogeneous corpora and controlled for level of homogeneity: mapping tag sets, token preprocessing, and tag normalization. The individual training sets performed better than the combined training set. However, models that were trained on the combined training set performed better. Data bias was found to be overcome by creating larger data sets.
Using C5.0 and Exhaustive Search for Boosting Frame-Semantic Parsing Accuracy – Guntis Barzdins, Didzis Gosko, Laura Rituma and Peteris Paikens
Barzdins et. al. use a decision tree classifier, manual rules, and exhaustive search to improve the accuracy of frame-semantic parsing, which is used in the FrameNet paradigm. They were able to achieve state-of-the-art results with a relatively small set of training data. The use case language was Latvian and an English parser was used as the standard for state-of-the-art performance. Furthermore, exhaustive search was used instead of the decision tree classifier to achieve enhanced results.
ML-Optimization of Ported Constraint Grammars – Eckhard Bick
Bick uses Machine Learning (ML) to optimize and port a manually written Constraint Grammar (CG) to another language. An English CG was ported to a Danish CG and achieved F-scores of 92.3. Techniques

of rule-promoting, frequency fail-safe, and word form relaxation contributed to the improved F-scores. In addition, optimization for a Danish CG led to error reduction of 10%.

A Deep Context Grammatical Model For Authorship Attribution – Simon Fuller, Phil Maguire and Philippe Moser

Fuller et. al. use a Probabilistic Context Free Grammar (PCFG) based variable-order Markov Model for authorship attribution. The model was compared to a linear kernel SVM and was found to have a slightly inferior performance at high sample sizes and a significantly inferior performance for smaller sample sizes.

Mapping Between English Strings and Reentrant Semantic Graphs – Fabienne Braune, Daniel Bauer and Kevin Knight

Braune et. al. compare formalisms of graph transduction, which is used to create semantic graphs from English strings. The formalisms compared were (1) synchronous hyperedge-replacement grammar (SHRG), (2) DAG-to-tree transducer (D2T), and (3) cascade of tree transducers (Cascade). The formalisms were tested for the tasks of natural language generation and understanding (NLG and NLU), each of which was evaluated using the BLEU and Smatch measures, respectively. SHRC and D2T required a higher number of rules than Cascade. The highest accuracy for NLG was achieved equally by D2T and Cascade. SHRC achieved highest NLU accuracy. The authors observed that the large number of rules required should be reduced.

Session O48 – Information Extraction and Text Structure (14:55-16:35)

Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web – Giuseppe Rizzo, Marieke van Erp and Raphael Troncy

Rizzo et. al. combine two approaches to named entities: NLP based recognition (NER) and semantic web based linking (NEL). They optimized them using a machine learning algorithm. They also employed NERD based extractors. They tested their approach on newswire data and microposts. The results showed that the approach yielded better performance. Some difficulties were attributed to a lack of consensus among various extractors regarding certain classes (person, location and organization) and also due to NEL being a fairly recent task.

Annotating Relations in Scientific Articles – Adam Meyers, Giancarlo Lee, Angus Grieve-Smith, Yifan He and Harriet Taber

Meyers et. al. annotated relations between entities in PubMed scientific articles. The motivation of this work is to improve methods of forecasting and prediction of trends in the field of technology. Relations between two entities or "arguments", are assigned signals, which can consist of words or grammatical labels (e.g., possessive, apposition, citation etc.). Evaluated relations were the following: EXEMPLIFY, ABBREVIATE, ORIGINATE, OPINION, and REL WORK. Only signals learnable via ML were employed. Pre-processing was conducted with manual rules and also based on previous annotation. Evaluation was carried out using a "strict" and a "sloppy" criteria. In the evaluation of the annotation, pre-processing, and post-processing, the "sloppy" criteria at times did not improve on the scores of the "strict" criteria; moreover, the improvement did not exceed 5% in F-scores. The project is in transition and post-processors will be built based on the pre-processors.

Improving Entity Linking using Surface Form Refinement – Eric Charton, Marie-Jean Meurs, Ludovic Jean-Louis and Michel Gagnon

Charton et. al., employ surface form generation and rewriting in an attempt to improve named entity resolution and linking. They employ an algorithm that suggests alternate surface forms; the system seeks to find alternate spellings or correct spellings. The methods yields improvement for all data types (web, forum, or news) and for all entity types (person, organization, or geopolitical entity). In the case

of entities not in the system's knowledge base, performance was inferior but not significantly.
Evaluating Improvised Hip Hop Lyrics - Challenges and Observations – Karteek Addanki and Dekai Wu
Addanki and Wu use inversion transduction grammars for the task of generating improvised hip hop lyrics. They carried out experiments for both the English and French languages. Disfluencies were dealt with either filtering or correction. The effectiveness of the two approaches was compared during the evaluation, which was conducted as 2-scale and 3-scale for the criteria of fluency and rhyming. The system was compared again an off-the-shelf phrase-based SMT system. Inter-evaluator agreement was improved with more examples and precise instructions. Inter-evaluator agreement was independent of the length of the data. Greater inter-evaluator agreement was found for the French data and it was explained as possible due to less ambiguity in the data. Language independence of inter-evaluator agreement will be further explored in the future.
Towards Automatic Detection of Narrative Structure – Jessica Ouyang and Kathy McKeown
Ouyang and McKeown use William Labov's theory of narrative analysis to create computational methods to detect narrative structure. They mapped elements of his structure with discourse relations in the Penn Discourse Treebank. They applied the approach to the detection of Complicating Actions with the result being an F-score of 71.55. They also achieved results that supported their hypothesis that the detection of narrative structure is related to the detection of discourse relations.

Closing Ceremony

The closing ceremony featured the presentation of the Antonio Zampoli prize and the celebration of the 15th anniversary of LREC. The winner of the Antonio Zampoli prize was Alex Waibel from Carnegie Mellon University (USA) and Karlsruhe Institute of Technology (Germany). The 15th anniversary celebration consisted of a presentation by Joseph Mariani in which he shared statistics and metrics about publications, authors, affiliations, and collaborations.

Post-Conference Workshop - Saturday, May 30, 2014

2nd Workshop on Language Resources and Evaluation for Religious Texts (LRE-Rel2)

Automatically-generated, phonemic Arabic-IPA Pronunciation Tiers for the Boundary-Annotated Qur'an Dataset for Machine Learning (version 2.0) – Majdi Sawalha, Claire Brierley and Eric Atwell
The plenary speaker was Majdi Swalha (University of Jordan and University of Leeds) who presented a paper he authored with Claire Brierley and Eric Atwell titled Automatically generated, phonemic Arabic-IPA pronunciation tiers for the Boundary Annotated Qur'an Dataset for Machine Learning (version 2.0). The research featured in the paper is centered around a corpus of the Quran (BAQ) whose boundaries are marked for specialized pronunciation or *tajwid*. It adds a grapheme to phoneme mapping from Standard Arabic to the International Phonetic Alphabet. This involves a two-phase algorithm since one-to-one mapping between Arabic letters and IPA entries is not sufficient. The second phase consists of rules that address issues of pronunciation such as special pronunciation of words or unpronounced vowels. Gold standard evaluation reached 100% accuracy. The algorithm was run for data in both Classical Arabic and Modern Standard Arabic and reached 99.55% and 99.48% accuracy, respectively. The BAQ was augmented with extra tiers including two which featured IPA transcriptions.
ABaC:us Revisited - Extracting and Linking Lexical Data from a Historical Corpus of Sacred Literature – Claudia Resch, Thierry Declerck, Barbara Krautgartner, and Ulrike Czeitschner
Thierry Declerck (DFKI, Germany) presented a paper titled ABaC:us revisited – Extracting and Linking Lexical Data from a historical Corpus of Sacred Literature. This paper was authored by himself along with Claudia Resch, Barbara Krautgartner, and Ulrike Czeitschner. The work presented was pertaining

to extracting lexical data from a corpus of sacred literature in historical German from the Baroque era, which was part of the larger ABaC:us project. A tool written for modern German was used for lemmatization and POS tagging. The difference between the two language variants along with orthographic variations in the data required the use of a specially developed tool to allow the output to be checked and corrected. Errors were reduced through the use of a corrected word list derived from a single document. The resulting lexicon contained associations of historical German lemma to modern German lemmas. This lexicon was also modeled in semantic web standards (RDF and SKOS) for the purpose of linking the word senses to those in the Linked Open Data (LOD) framework.

Combining Critical Discourse Analysis and NLP Tools in Investigations of Religious Prose – Bruno Bisceglia, Rita Calabrese, Ljubica Leone

Rita Calabrese presented a paper she authored with Ljubica Leone and Bruno Bisceglia (University of Salerno) entitled Combining Critical Discourse Analysis and NLP tools in investigations of religious prose. In this work, the authors analyzed data from the First and Second Vatican Councils. The texts were tagged with VISL parsers. Critical Discourse Analysis (CDA) was employed and the focused was on auxiliary verbs (can, could, may, might, must, shall, should, will, and would). The raw frequency (Rf) and normalized frequency (Nf) was computed for declarations, constitution, and decree. "Must" was the most frequent auxiliary in the constitution category whereas "should" was the most frequent in the other two categories. The high frequency of "should" was found to be due to it being less face threatening. Furthermore, intrinsic vs. extrinsic modalities were considered. Extrinsitcs were of low frequency, which led to the conclusion that subcorpora featuring it show respect for human willingness. The use of NLP tools and CDA within religious discourse in this study helped the researchers gain insight into the attitudes of theologians towards their audience.

Humour and Non-Humour in Religious Discourse – Daniela Gfu, Liviu-Andrei Scutelnicu and Dan Cristea

Dan Cristea (Alexandru Ioan Cuza University; Romanian Academy) presented a paper titled Humour and non-humour in religious discourse. This work was a result of a collaboration with Daniela Gifu (Alexandru Ioan Cuza University) and Liviu-Andrei Scutelnicu (Romanian Academy). Based on the observation that religious texts, namely sermon, contain humorous features, the authors conducted a pilot study to identify such features. The study focused on Adjectival Noun Phrases (ANPs). The data was in Romanian and contained the sermons of the monk priest Ilie Cleopa. A religious lexicon containing a hierarchy of synsets was created; it contained 2,367 entries. Humorous ANPs (HANP) were manually annotated as religious (RHANP) and non-religious (NRHANP). Similar phrases were also automatically detected using lexical pattern rules (e.g., noun + article + adjective). Religious humorous ANPs were identified with precision, recall, and f-measures of over 80%. The non-religious counterparts were identified with similar precision, but with lower recall and f-measure of a little over 60% and 70%, respectively.

A Proposed Model for Qur'anic Arabic Wordnet – Manal AlMaayah, Majdi Sawalha and Mohammad A.M. Abushariah

Majdi Sawalha (University of Jordan) presented a paper titled A Proposed Model for Quranic Arabic WordNet which was co-authored with Manal AlMaayah Mohammad A. M. Abushariah. This work seeks to create a Quranic Classical Arabi WordNet, seeing that most Arabic NLP research favors MSA and to a great extent neglects classical Arabic. Available resources in this endeavor are (1) the previously mentioned BAQ corpus which provides word roots, POS, and English translations, (2) standard Arabic dictionaries to find words derived for each root, and (3) Arabic WordNet and Quran Ontology for establishing semantic relations of synonym, antonym, and similarity.This work is presents a proposed model and the details of the implementation are a work in progress.

The Greek-Arabic New Testament Interlinear Process: greekarabicnt.org – Kamal Abou Mikhael

Kamal Abou Mikhael (American University of Beirut), the author of this report, presented a paper titled The Greek-Arabic New Testament Interlinear Process: greekarabicnt.org. This work presents a methodology for creating Greek-Arabic New Testament interlinears, which have yet to be published in a digital text format. The first output of such a process is a text-based aligned corpus. This is achieved using standard nix commands, tools, and scripting languages as well as freely available linguistic tools and NT texts. Aligning Greek and Arabic requires the selective segmentation of Arabic texts that takes into consideration the morphological correspondence between the two language, namely how certain bound morphemes in Arabic correspond to free morphemes in Greeks. A database of segmentation solutions was developed and used to segment the Arabic text. The Greek and Arabic texts, which were converted to a word-per-line format, were manually aligned using the *vim* editor. The first two chapters of the Epistles of John were aligned and displayed in interlinear and reverse-interlinear form using *LaTeX*. An average of 13% of the words needed re-ordering and most of the alignment work consisted of inserting gaps to account for untranslated text or translational additions.

The oral presentations in the workshop were followed by poster presentations. The workshop was concluded with a discussion regarding how the greater interest can be generated in the workshop. It was mentioned that there was an attempt to include more participants by allowing historical texts and non-standard religious discourse (such as myths, etc.) to be included. The following were some suggestions: (1) that there be attempts to apply methodologies already presented for certain texts to other texts (e.g., interlinears for the Bible attempted for the Quran or pronunciation markup in the Quran applied to the Bible) and (2) that there be shared tasks.In addition to workshop participation, it was stated that the workshop had the aim of fostering inter-faith corpus studies, exchange of ideas regarding the processing of religious texts, and generating interests among theologians in the use of NLP tools and methods to explore religious text.