Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools

Workshop Programme

Date 27 May 2014

09:00 – 10:30 Session 1

09:00 – 09:20 Welcome and Introduction by Workshop Chair

09:20 – 09:50 Keynote Speech by Prof. Mansour Algamdi
King Abdullah Initiative for Arabic Content

09:50 – 10:10 Wajdi Zaghouani
Critical Survey of the Freely Available Arabic Corpora

10:10 -10:30 Jonathan Forsyth
Automatic Readability Prediction for Modern Standard Arabic

10:30 – 11:00 Coffee break

11:00 -13:00 Session 2

11:00 – 11:20 Eshrag Refaee and Verena Rieser
Subjectivity and Sentiment Analysis Of Arabic Twitter Feeds With Limited Resources

11:20 – 11:40 Ali Meftah, Yousef Alotaibi and Sid-Ahmed Selouani
Designing, Building, and Analyzing an Arabic Speech Emotional Corpus

11:40 – 12:00 Thomas Eckart, Uwe Quasthoff, Faisal Alshargi and Dirk Goldhahn
Large Arabic Web Corpora of High Quality: The Dimensions Time and Origin

12:00 – 12:20 Ryan Cotterell and Chris Callison-Burch
An Algerian Arabic / French Code-Switched Corpus

12:20:12:40 Mourad Loukam, Amar Balla and Mohamed Tayeb Laskri
An Open Platform Based on HPSG Formalism for the Standard Arabic Language

12:40 – 13:00 Ghania Droua-Hamdani, Yousef Alotaibi, Sid-Ahmed Selouani and Malika Boudraa
Rhythmic Features Across Modern Standard Arabic and Arabic Dialects

Editors

Hend S. Al-Khalifa	King Saud University, KSA
Abdulmohsen Al-Thubaity	King AbdulaAziz City for Science and Technology, KSA

Workshop Organizers/Organizing Committee

Hend S. Al-Khalifa	King Saud University, KSA
Abdulmohsen Al-Thubaity	King AbdulaAziz City for Science and Technology, KSA

Workshop Programme Committee

Eric Atwell	University of Leeds, UK
Khaled Shaalan	The British University in Dubai (BUiD), UAE
Dilworth Parkinson	Brigham Young University, USA
Nizar Habash	Columbia University, USA
Khurshid Ahmad	Trinity College Dublin, Ireland
Abdulmalik AlSalman	King Saud University, KSA
Maha Alrabiah	King Saud University, KSA
Saleh Alosaimi	Imam University, KSA
Sultan almujaiwel	King Saud University , KSA
Adam Kilgarriff	Lexical Computing Ltd, UK
Amal AlSaif	Imam University, KSA
Maha AlYahya	King Saud University ,KSA
Auhood AlFaries	King Saud University , KSA
Salwa Hamadah	Taibah University, KSA
Abdullah Alfaifi	University of Leeds, UK

Table of contents

Author Index

Preface

For Natural Language Processing (NLP) and Computational Linguistics (CL) communities, it was a known situation that Arabic is a resource poor language. This situation was thought to be the reason why there is a lack of corpus based studies in Arabic. However, the last years witnessed the emergence of new considerably free Arabic corpora and in lesser extent Arabic corpora processing tools.

This workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) aimed to encourage researchers and developers to foster the utilization of freely available Arabic corpora and open source Arabic corpora processing tools and help in highlighting the drawbacks of these resources and discuss techniques and approaches on how to improve them.

OSACT had an acceptance rate of 67%, we received 12 papers from which 8 papers were accepted. We believe the accepted papers are high quality and present mixture of interesting topics. We would like to thank all people who in one way or another helped in making this workshop a success. Our special thanks go to Professor Mansour Alghamdi for accepting to give the invited presentation, to the members of the program committee who did an excellent job in reviewing the submitted papers, to Saad Alotaibi for designing and updating OSACT website and to the LREC organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Hend Al-Khalifa and Abdulmohsen Al-Thubaity
Reykjavik (Iceland), 2014

Critical Survey of the Freely Available Arabic Corpora

Wajdi Zaghouani
Carnegie Mellon University Qatar
Computer Science
E-mail: wajdiz@cmu.edu

Abstract

The availability of corpora is a major factor in building natural language processing applications. However, the costs of acquiring corpora can prevent some researchers from going further in their endeavours. The ease of access to freely available corpora is urgent needed in the NLP research community especially for language such as Arabic. Currently, there is not easy was to access to a comprehensive and updated list of freely available Arabic corpora. We present in this paper, the results of a recent survey conducted to identify the list of the freely available Arabic corpora and language resources. Our preliminary results showed an initial list of 66 sources. We presents our findings in the various categories studied and we provided the direct links to get the data when possible.

Keywords: Arabic, Open source, Free, Corpora, Corpus, Survey.

1. Introduction

The use of corpora has been a major factor in the recent advance in natural language processing development and evaluation. However, the high costs of building or licensing a corpora could be an obstacle for many young researchers or even some institution in several parts of the world. Therefore, having access to freely available corpora is clearly a desirable goal. Unfortunately, the freely available corpora are generally not easily found and most resources available from language data providers are for fees or exclusively reserved for subscribers, such as the corpora available from the Linguistic Data Consortium or the Evaluations and Language resources Distribution Agency (ELDA). A simple query for Arabic corpora available in the LDC Catalog shows the availability of 116 corpora of various types (text, speech, evaluation etc…).[1] Another similar query done with the ELRA Corpora search engine showed the availability of 80 corpora. [2] For instance, Arabic can still be considered a relatively resource poor language when compared to other languages such as English, and having access to freely available corpora will definitely improve the current Arabic NLP technologies. In this paper, we present the results of an online survey of the freely available Arabic corpora.

2. Current situation of the freely available Arabic corpora

Before starting our survey experiment, we tried various online queries to locate any freely available Arabic corpora or a repository listing the corpora for any easy access to the resources. We found that the information is scattered in various personal and research groups sites that are often not complete or outdated.

As of 2010, ELRA created the LRE Map (Language Resources and Evaluation) which is an online database on language resources. The goal behind LRE Map is to monitor the creation and the use and of language resources. The information is collected during the submission process to LREC and other conferences. We did a query to list the freely available Arabic corpora and we found a limited number and no URL to link the user or project details were available. Habash (2010) listed in his book, various available Arabic corpora sorted by corpus type (speech, text, bi-lingual). Again the list is not designed for the freely available resources and most of data listed are available from data providers. The Association for Computational Linguistics (ACL) maintains a wiki page that lists the available resources by language, the Arabic page only lists five corpora, four free corpora and one proprietary corpora.[3]

The European Network of Excellence in Human Language Technologies (ELSNET) maintains a list of pointers to Arabic and other Semitic NLP and Speech sites, the Arabic resources section includes 23 entries and most of them were created more than 12 years ago.[4]

The Mediterranean Arabic Language and Speech Technology (MEDAR) conducted a survey in 2009 and 2010 to list the existing institutions and experts involved in the development of Arabic language resources, activities and projects being carried out and related and tools.[5] The collected results were compiled and made available into a knowledge base that is accessible online[6]. Again, despite the huge effort made, the list is no

[1] Query performed on January 31st 2014 <http://catalog.ldc.upenn.edu/>

[2] Query performed on January 31st 2014 <http://catalog.elra.info/search.php>

[3] http://aclweb.org/aclwiki/index.php?title=Resources_for_Arabic

[4] http://www.elsnet.org/arabiclist.html

[5] http://www.medar.info/MEDAR_Survey_III.pdf

[6] www.elda.org/medar_knowledge_base/

longer updated and it lacks the necessary information to locate the data such as the download page or the project description and URL. Finally, we cite other interesting personal efforts to list some Arabic language resources such as the Sibawayh Repository for Arabic Language Processing page,[7] Al-Sulaiti Arabic corpora page[8] and Al-Ghamidi Arabic links page. [9] We consider our efforts described in this project as a complement to what exists already with a focus on the free resources and how each corpus can be obtained.

3. The Survey

In order to start the collection of our freely available Arabic corpora list, we created an online survey[10] that was shared in the various NLP related lists such as Corpora[11] and ArabicList.[12] The online survey was intended to be completed within 5-10 minutes to encourage participants, and included some very basic questions such as the provider information, corpus type, size, purpose, download link, related publications, Arabic variety, production status and a confirmation that the corpus is completely free for a research purposes. The online survey was completed by 20 participants who pointed 26 resources. Once the survey results were compiled, we added manually, 40 other freely available Arabic resources taken from the various online sources described in section 2. We also added the missing information when needed (wrong download URL, description, corpus size, authors, etc.). Finally we tried to locate any related publication to the source so it can be cited properly when used. In the next section, we describe briefly a selection of the 66 free resources found during our survey.

4. Available Resources

In this section, we present the result of our survey of the freely available Arabic corpora with a focus on the most important work for each of the following categories:

- **Raw Text Corpora**: monolingual corpora, multilingual corpora, dialectal Corpora, web-based corpora.
- **Annotated Corpora:** named entities, error annotation, POS, syntax, semantic, anaphora.
- **Lexicon:** lexical databases and words lists.
- **Speech Corpora:** audio recording, transcribed data.
- **Handwriting Recognition Corpora:** scanned and annotated documents**.**
- **Miscellaneous Corpora types:** Questions/Answers, comparable corpora, plagiarism detection and summaries.

For each of the four categories, some basic information will be provided a table for each category. It includes the author name or the research group, the corpus name, the corpus size in words or in files. In case there is a publication associated with the corpus, it will be cited as part of the author name in the table otherwise only the corpus download/access URL is provided as a footnote following the corpus name in each table. The sources are sorted in the tables according to their size for the most important to the least important.

4.1 Raw Text Corpora

In this section we cite 23 freely available raw text corpora, that is, they do not include any kind of annotation and limited to the text files themselves. The raw text corpora are divided into four categories listed below.

4.1.1. Monolingual Corpora

The 11 freely available monolingual corpora found are all available for download (Table 1).

Source	Corpus	Words
Abdelali	Ajdir Corpora[13]	113,000,000
Alrabiah	KSU Corpus of Classical Arabic[14]	50,000,000
Saad and Ashour (2010)	OSAC[15]	18,183,511
Abbas	Alwatan[16]	10,000,000
Zarrouki	Tashkeela[17]	6,149,726
Abbas	Al Khaleej[18]	3,000,000
Al-Thubaity et al.	KACST Arabic Newspaper Corpus[19]	2,000,000
Al-Saadi	Arabic Words Corpora[20]	1,500,000
Al-Suleiti	Corpus of Contemporary Arabic[21]	842,684
Alkanhal et al. (2012)	CRI KACST Arabic Corpus[22]	235,000
Farwaneh	Arabic Learners Written Corpus[23]	50,000

Table 1: Monolingual Corpora List.

Most of them cover the news domain and they are large size corpora ranging from 1 million words to 113 million words. Other corpora cover other domains such as student essays (Farawaneh) and classical Arabic (KSU Corpus of classical Arabic and Tashkeela). When it comes to data format, we noticed that most of these corpora are stored in

[7] http://arab.univ.ma/web/s/

[8] http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm

[9] http://www.mghamdi.com/links.htm

[10] https://docs.google.com/forms/d/1N2W76d8Uxnzx--0Dj6An2mJr8KzeR0U1rF6pOj6Djjg/viewform?edit_requested=true

[11] http://www.hit.uib.no/corpora/

[12] https://listserv.byu.edu/cgi-bin/wa?A0=ARABIC-L

[13] http://aracorpus.e3rab.com/argistestsrv.nmsu.edu/AraCorpus/

[14] http://ksucorpus.ksu.edu.sa/?p=43

[15] https://sites.google.com/site/motazsite/arabic/osac

[16] http://sourceforge.net/projects/arabiccorpus/

[17] http://sourceforge.net/projects/tashkeela/

[18] http://sourceforge.net/projects/arabiccorpus/

[19] http://sourceforge.net/projects/kacst-acptool/files/?source=navbar

[20] http://sourceforge.net/projects/arabicwordcorpu/files/

[21] http://www.comp.leeds.ac.uk/eric/latifa/research.htm

[22] http://cri.kacst.edu.sa/Resources/TRN_DB.rar

[23] http://l2arabiccorpus.cercll.arizona.edu/?q=allFiles

text or xml format while others like the Arabic learners corpus, are stored in an inconvenient PDF format that makes it hard to be used for any NLP task.

4.1.2. Multilingual Corpora

Among the corpora listed in Table 2, we can consider the UN corpus as the most important and the most widely known free corpus for its category. The Meedan with 1 million words Arabic/English aligned sentences, is also a very valuable resource. The Hadith standard corpus and the Quranic Arabic/English aligned corpus included in the Egypt translation tool are less known resources that could be used in any work related to the religious domain.

Source	Corpus	Words
Rafalovitch and Dale (2009)	UN Corpus(Arabic portion)[24]	2,721,463
Bounhas	Hadith Standard Corpus[25]	2,500,000
Meedan	MEEDAN Translation Memory[26]	1,000,000
CLSP/JHU	EGYPT Translation Toolkit[27]	80,000

Table 2: Multilingual Corpora List.

4.1.3. Dialectal Corpora

The two dialectal corpora listed in the Table 3 below are very valuable, especially that work Arabic dialect processing is a rather recent task there is a real need for such resources. The Tunisian Dialect Corpus (Graja et al.), is a transcribed spoken dialogue corpus formed of 1465 railway staff utterances and 1615 client utterances. The recent work done by (Almeman and Lee 2013), can be considered a major contribution to the advance in the Arabic dialectal resources with its 2 million unique words collected online from 55k webpages and covering four major Arabic dialects (Gulf, Levantine, North Africa, Egypt).

Source	Corpus	Words
Almeman and Lee (2013)	Arabic Multi Dialect Text Corpora[28]	2,000,000
Graja et al. (2010)	Tunisian Dialect Corpus (TuDiCoI)[29]	3,403

Table 3: Caption.

4.1.4. Web-based Corpora

In this category we placed some corpora (Table 4) that are exclusively available online through an online query interface so there is no data provided for download which can be inconvenient for some research studies, nevertheless these web-based corpora can be very valuable for concordance and frequency studies given the variety and large size of these corpora (732M words KACST corpus, 317M M words for Leeds and 100M words for ICA and Parkinson corpus), moreover the Arabic variety and text genre covered is large which makes these corpora very suitable for all types of Arabic linguistics studies (Quranic Arabic, classic Arabic, newswire, books etc.).

Source	Corpus	Words
Al-Thubaity	KACST Arabic Corpus[30]	732,780,509
Leeds	Leeds Arabic Internet Corpus[31]	317,000,000
Alansary et al. (2007)	International Corpus of Arabic[32]	100,000,000
Parkinson	ArabiCorpus[33]	100,000,000
Abbas N.	QURANY[34]	78,000
Sharaf et al.	Quranic Text mining Dataset[35]	24,000

Table 4: Web-based Corpora List.

4.2 Annotated Corpora

Annotated corpora are very useful to build systems and tools based on supervised algorithms and the free availability of resources will help young researches to train and build systems at minimal cost. In this section, we list a selection of freely available named entities corpora, Error annotated corpora and some various annotated corpora including part of speech (POS) annotated corpora, syntactically and semantically annotated corpora.

4.2.1. Named Entity Corpora

Table 5 lists some very useful resources for the named entities recognition task.

Source	Corpus	Words
Steinberger et al. (2011)	JRC-Names[36]	230,000
Ben Ajiba et al. (2007)	ANERCorp[37]	150,000
Mohit et al. (2012)	AQMAR Named Entity Corpus[38]	74,000
Azab et al. (2013)	Named Entity Translation Lexicon[39]	55,000
Attia et al. 2010	Named Entities List[40]	45,202
Ben Ajiba et al. (2007)	ANERGazet[41]	14,000

Table 5: Named Entities Corpora List.

Most of these corpora were reported by their respective authors in major NLP conferences which adds visibility to these resources. The annotation format of these data

[24] http://www.uncorpora.org/

[25] http://www.kunuz/

[26] https://github.com/anastaw/Meedan-Memory

[27] http://old-site.clsp.jhu.edu/ws99/projects/mt/toolkit/

[28] http://www.cs.bham.ac.uk/~kaa846/arabic-multi-dialect-text-corpora.html

[29] https://sites.google.com/site/anlprg/outils-et-corpus-realises/TuDiCoIV1.xml?attredirects=0

[30] http://www.kacstac.org.sa/

[31] http://smlc09.leeds.ac.uk/query-ar.html

[32] http://www.bibalex.org/ica/en/About.aspx

[33] http://arabicorpus.byu.edu/

[34] http://quranytopics.appspot.com/

[35] http://textminingthequran.com/wiki/Main_Page

[36] http://ipsc.jrc.ec.europa.eu/index.php?id=42#c2696

[37] http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

[38] http://www.ark.cs.cmu.edu/ArabicNER/

[39] http://nlp.qatar.cmu.edu/resources/NETLexicon/

[40] https://sourceforge.net/projects/arabicnes/

[41] http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

follows the XML annotation standards put by major evaluation campaigns such as the automatic content extraction (ACE) evaluation campaign.[42] Most of the entries in these resources covers person's names, some organisations and geographical locations names and the size of these data is important, ranging from 14k to 230k.

4.2.2. Error-Annotated Corpora

Error annotated corpora can be very useful for corpus based studies of errors and also for building automatic spelling correction tools. Table 6 lists three resources, QALB and the Arabic learner corpus are still an on-going efforts. The KACST Error corpus (Alkanhal et al. 2012) includes exclusively student essays that are manually corrected while Alfifi et al. (2013) will include the correction of the errors as well as the categories of the errors. When ready, Zaghouani et al. (2014) corpus will be the only 2M words corrected corpus available for Arabic with four text varieties: native and non-native students essays, online users posts and English/Arabic machine translation corrected output.

Source	Corpus	Words
Habash et al. (2013)	Qatar Arabic language Bank(QALB)[43]	2,000,000
Alfifi et al. (2013)	Arabic Learner Corpus[44]	282,000
Alkanhal et al. (2012)	KACST Error Corrected Corpus[45]	65,000

Table 6: Errors Annotated Corpora List

4.2.3. Miscellaneous Annotated Corpora

The corpora listed on Table 7 includes various annotated corpora ranging from semantically annotated corpora to syntactically and morphologically annotated corpora.
Most of these resources allows direct download except for OntoNotes that can be obtained freely from the LDC.

The OntoNotes corpus (Weischedel et al. 2013) includes various genres of text (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows) in three languages (English, Chinese, and Arabic) with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference).

One notable effort in this category is the ongoing work to build the Quranic Arabic Corpus, an annotated linguistic resource consisting of 77,430 words of Quranic Arabic. The project aims to provide morphological and syntactic annotations for researchers wanting to study the language of the Quran. While the POS annotated version is already available for download, the treebank version is still ongoing. Moreover, an online query interface is available for morphological queries and concordance. Despite its small size, the Arabic Wikipedia dependency corpus is one of the rare freely available Arabic Treebanks.

Source	Corpus	Words	Type
Weischedel et al. (2013)	OntoNotes Release 5.0[46]	300,000	Semantic
Dukes and Habash (2010)	The Quranic Arabic Corpus[47]	77,430	POS/Syntax
Schneider et al. (2012)	AQMAR Arabic Wiki. Supersense Corpus[48]	65,000	Semantic
Khoja et al. (2001)	Khoja POS tagged corpus[49]	51,700	POS
E. Mohammed	Arabic Wikipedia Dependency Corpus[50]	36,000	Syntax
Mezghani et al. (2009)	AnATAr Corpus[51]	18,895	Anaphora

Table 7: Miscellaneous Annotated Corpora List.

4.3 Lexicon

In this section we describe some available lexical databases and words lists. Most of these resources are available for download, some of the lexicon are part of tools and systems, but since these tools are open source, these lexicons can be used for research purposes.

4.3.1. Lexical Databases

Several efforts have been made in recent years to build various lexical resources for Arabic (Table 8). Fortunately, most of them are free such as the version 1.0 of the well-known Buckwalter morphological analyzer (Buckwalter 2002). Other important efforts were adapted from the English to the Arabic such as the Arabic WordNet (Elkateb et al. 2006) and the Arabic VerbNet (Mousser 2010).

In the Arabic WordNet, the words are grouped into sets of synonyms and it provides general definitions and the various semantic relations between the synonyms sets.
The Arabic VerbNet provides a lexicon in which the most used Arabic verbs are classified and their syntactic and semantic information are provided. An online interface is provided.

4.3.2. Words Lists

Table 9 lists various words lists created mostly by Mohammed Attia.[52] These words lists can be used by lexicographers to study various aspects of the Arabic language such as the Arabic MSA word count list. These

[42] http://www.itl.nist.gov/iad/894.01/tests/ace/

[43] http://nlp.qatar.cmu.edu/qalb/

[44] http://www.comp.leeds.ac.uk/scayga/alc/corpus%20files.html

[45] http://cri.kacst.edu.sa/Resources/TST_DB.rar

[46] http://catalog.ldc.upenn.edu/LDC2013T19

[47] http://corpus.quran.com/download/

[48] http://www.ark.cs.cmu.edu/ArabicSST/

[49] http://zeus.cs.pacificu.edu/shereen/research.htm#corpora and email the author

[50] http://www.ark.cs.cmu.edu/ArabicDeps/

[51] https://sites.google.com/site/anlprg/outils-et-corpus-realises/AnATArcorpus-BEB.rar?attredirects=0

[52] http://www.attiaspace.com/

lists can also be integrated with the lexicons of systems and tools to improve their performances. For instance, the Arabic wordlist of 9M words and the 18k Arabic unknown words list, can be used in a the spell checking systems. Furthermore, the Arabic stop words list of 13k can be used in various application as word filter list.

Source	Corpus	Words
Buckwalter	BAMA 1.0 English-Arabic Lexicon[53]	82,158
Salmone	Arabic-English Learner's Dictionary[54]	74,000
Doumi et al. (2013)	Unitex Arabic Package[55]	50,407
Boudelaa and Wilson (2010)	ARALEX Online[56]	37,494
Attia et al. (2011)	AraComLex Arabic Lexical Database[57]	30,000
Mousser (2010)	Arabic VerbNEt[58]	23,341
Elkateb et al. (2006)	Arabic WordNet[59]	18,957
Mesfar and Silberztein (2008)	NOOJ Arabic Dictionary[60]	10,000
ArabEyes	Qamoose[61]	N.A

Table 8: Lexical Databases List.

Source	Corpus	Words
Attia et al. (2011)	Word Count of Modern Standard Arabic[62]	1,000,000,000
Attia et al. (2012a)	Arabic Wordlist for Spellchecking[63]	9,000,000
Attia et al. (2010)	Multiword Expressions[64]	34,658
Attia et al. 2012b	Arabic Unknown Words[65]	18,000
Zarrouki	Arabic Stop words[66]	13,000
Attia et al. 2011b	Obsolete Arabic Words[67]	8,400
Attia et al. 2011c	Arabic Broken Plurals[68]	2,562

Table 9: List of Words Lists.

4.4 Speech Corpora

To the best of our knowledge the corpus in Table 10 compiled by Almeman and Lee (2013) is the only freely available speech corpus for Arabic. Most of the currently available speech corpora are available from the LDC or ELRA with a membership fees.

Source	Corpus	Files
Almeman and lee (2013)	Arabic Speech Corpora[69]	67,132

Table 10: List of Speech Corpora.

4.5 Handwriting Recognition Corpora

Again, the handwriting recognition corpora are very rare in Arabic and they are mostly available at cost. The four corpora listed in Table 11 are an exception and they can be used for various NLP tasks from OCR to writer identification.

Source	Corpus	Files
Al-Maadeed, et al (2011)	QUWI Handwritings Dataset[70]	1,000
Hassaïne and Maadeed (2012)	Writer Identification Contest for Arabic Scripts Data set[71]	200
Al-Maadeed, et al (2002)	AHDB Data Set[72]	100
Al-Maadeed, et al (2012)	ICDAR2011 competition Data set[73]	50

Table 11: Handwriting Recognition Corpora.

4.6 Miscellaneous Corpora types

The list in Table 12 presents seven corpora useful for a multitude of NLP related tasks such as question answering Ben Ajiba et al. (2007) and Trigui et al. (2010), plagiarism detection Bensalem et al. (2013), document summarization El-Haj et al. 2010 and El-Haj and Rayson (2013), comparable text detection Saad et al. (2013).

Finally, the Kalimat multi-purpose corpus (El-Haj and Koulali (2013) is a unique corpus that includes around 20k newswire words extracted for summaries, named entities tagged, part of speech tagged and morphologically analyzed.

5. Conclusion

We presented the preliminary results of the first survey reserved for the freely Arabic Corpora. The goal behind this study is to promote the use of free corpora especially by those who lack funding and cannot afford membership or high fees to acquire a corpora from a language data center. The results obtained showed that many of the freely available resources for Arabic are not always visible and therefore it is hard be found by potential users.

[53] http://catalog.ldc.upenn.edu/LDC2002L49

[54] http://www.perseus.tufts.edu/hopper/opensource/download

[55] http://www-igm.univ-mlv.fr/~unitex/index.php?page=3&htm

[56] https://aralex.mrc-cbu.cam.ac.uk/aralex.online/login.jsp

[57] http://sourceforge.net/projects/aracomlex/files/

[58] http://ling.uni-konstanz.de/pages/home/mousser/files/Arabic_verbnet.php

[59] http://sourceforge.net/projects/awnbrowser/

[60] http://www.nooj4nlp.net/pages/arabic.html

[61] http://sourceforge.net/projects/arabeyes/files/QaMoose/2.1/

[62] http://arabicwordcount.sourceforge.net/

[63] http://sourceforge.net/projects/arabic-wordlist/

[64] https://sourceforge.net/projects/arabicmwes/

[65] http://arabic-unknowns.sourceforge.net/

[66] http://sourceforge.net/projects/arabicstopwords/

[67] http://obsoletearabic.sourceforge.net/

[68] http://broken-plurals.sourceforge.net/

[69] http://www.cs.bham.ac.uk/~kaa846/arabic-speech-corpora.html

[70] http://handwriting.qu.edu.qa/dataset/

[71] http://handwriting.qu.edu.qa/dataset/

[72] http://handwriting.qu.edu.qa/dataset/

[73] http://handwriting.qu.edu.qa/dataset/

Moreover, the 66 corpora listed in this paper cover the main categories of corpora types. We hope that this initial attempt to located freely available Arabic corpora would be useful to the research community and such effort can be easily replicated to located similar sources for other languages. The corpora list presented in this paper, is made available in a single webpage for an easier access.[74] In the near future, we plan to make the list available in an online database and we will continue looking for other free corpora to enrich our repository.

Source	Corpus	Words
Saad et al. (2013)	AFEWC and Enews Comparable Corpora[75]	28,000,000
Bensalem et al. (2013)	InAra (a corpus for Arabic Intrinsic plagiarism detection evaluation)[76]	12,681,374
El-Haj et al. (2010)	Essex Arabic Summaries Corpus[77]	41,493
El-Haj and Rayson (2013)	Multi-document Summaries[78]	30,000
El-Haj and Koulali (2013)	KALIMAT Multi-Purpose Corpus[79]	20,291
Ben Ajiba et al. (2007)	Arabic QA/IR[80]	11,638
Trigui et al. (2010)	Arabic Definition QA corpus[81]	250

Table 12: Miscellaneous Corpora Types

6. References

Alansary, S., Nagi, M. and Adly, N. (2007). Building an International Corpus of Arabic (ICA): progress of compilation stage. In Proceedings of the 7th International Conference on Language Engineering, Cairo, Egypt, 5–6 December 2007.

Alfaifi, A, Atwell, E and Abuhakema, G. (2013). Error Annotation of the Arabic Learner Corpus: A New Error Tagset. In Language Processing and Knowledge in the Web, Lecture Notes in Computer Science. 25th International Conference, GSCL 2013, 25-27 September 2013, Darmstadt, Germany. Springer , 14-22 (9).

Alkanhal, M.I., Al-Badrashiny, M.A., Alghamdi, M.M., Al-Qabbany, A.O. (2012). Automatic Stochastic Arabic Spelling Correction With Emphasis on Space Insertions and Deletions. In IEEE Transactions on Audio, Speech, and Language Processing. 20(7): 2111-2122, Sept. 2012.

Alkanhal, M. I., Al-Badrashiny, M.A., Alghamdi, M. M., Al-Qabbany, Abdulaziz, O. (2012). Automatic Stochastic Arabic Spelling Correction With Emphasis on Space Insertions and Deletions. IEEE Transactions on Audio, Speech and Language Processing 20(7): 2111-2122.

Al-Maadeed, S., D. Elliman, C.A. Higgins. (2002). A database for Arabic handwritten text recognition research. In Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition. Niagara-on-the-Lake, Canada, 2002.

Al-Maadeed, S., W. Ayouby, A. Hassaïne, J. Alja'am. (2012). QUWI: An Arabic and English Handwriting Dataset for Offline Writer Identification. In Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition, Bari, Italy, pp. 742-747.

Almeman K., M. Lee and A. Almiman. (2013). Multi Dialect Arabic Speech Parallel Corpora, In Proceedings of the First International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13), Sharjah, UAE, 12-14 Feb. 2013.

Almeman K. and M. Lee. (2013). Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words, In Proceedings of The First International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13), Sharjah, UAE, 12-14 Feb. 2013.

Alrabiah, M., Al-Salman, A. and Atwell, E. (2013). The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic. In Proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL-2), 22 Jul. 2013, Lancaster University, UK.

Al-Thubaity, A., Khan, M., Al-Mazrua, M., Al-Mousa, M. (2013). New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool. In Proceedings of the Asian Language Processing (IALP) Conference, pp.67,70.

Attia, M. Pecina, P., Tounsi, L., Toral, A., Van Genabith, J. (2011). A Lexical Database for Modern Standard Arabic Interoperable with a Finite State Morphological Transducer. In Mahlow, Cerstin; Piotrowski, Michael (Eds.) Systems and Frameworks for Computational Morphology. Second International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011.

Attia, M., Toral, A., Tounsi, L., Monachini, M. and Van Genabith, J. (2010). An automatically built Named Entity lexicon for Arabic. In Proceedings of the Language Resources and Evaluation Conference (LREC) 2010. Valletta, Malta.

Attia, M., Toral, A., Tounsi, L., Pecina, P. and Van Genabith, J. (2010). Automatic Extraction of Arabic Multiword Expressions. In Proceedings of the COLING Workshop on Multiword Expressions: from Theory to Applications. Beijing, China.

Attia, M., Pecina, P., Tounsi, L., Toral, A., Van Genabith, J. (2011b). A Lexical Database for Modern Standard Arabic Interoperable with a Finite State Morphological Transducer. In Mahlow, Cerstin; Piotrowski, Michael (Eds.) Systems and Frameworks for Computational Morphology. Second International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011, Proceedings. Series: Communications in Computer and Information Science, Vol. 100. 1st Edition.

Attia, M., Pecina, P., Tounsi, L., Toral, A. and Van Genabith, J. (2011c). Lexical Profiling for Arabic.

[74] The current list of corpora is available at <http://www.qatar.cmu.edu/~wajdiz/corpora.html>

[75] http://sourceforge.net/projects/crlcl/

[76] https://sourceforge.net/projects/inaracorpus/

[77] http://sourceforge.net/projects/easc-corpus/

[78] http://multiling.iit.demokritos.gr/file/all

[79] http://sourceforge.net/projects/kalimat/

[80] http://www.dsic.upv.es/~ybenajiba/resources/Corpus.zip

[81] https://sites.google.com/site/anlprg/outils-et-corpus-realises/ArabicDefinitionQuestionAnsweringData.rar?attredirects=0

Electronic Lexicography in the 21st Century. Bled, Slovenia.

Attia, M., Pecina, P., Tounsi, L., Toral, A. and Van Genabith, J. (2011). A Lexical Database for Modern Standard Arabic.

Attia, M., Pecina, P., Samih, Y., Shaalan, K., Van Genabith, J. (2012). Improved Spelling Error Detection and Correction for Arabic. In Proceedings of the COLING 2012, Bumbai, India.

Attia, M., Samih, Y., Shaalan, K., Van Genabith, J. (2012b). The Floating Arabic Dictionary: An Automatic Method for Updating a Lexical Database. In Proceedings of the COLING 2012, Bumbai, India.

Azab M. , Bouamor, H., Mohit, B. and Oflazer, K. (2013). Dudley North visits North London: Learning When to Transliterate to Arabic. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT 2013), Atlanta, USA, June 2013.

Benajiba Y., Rosso, P. and BenediRuiz, J-M. (2007). Anersys: An Arabic named entity recognition system based on maximum entropy. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 4394 of Lecture Notes in Computer Science, pages 143–153. Springer Berlin / Heidelberg.

Benajiba, Y., Rosso, P., Soriano, JMG. (2007). Adapting the JIRS Passage Retrieval System to the Arabic Language. In Proceedings of the Computational Linguistics and Intelligent Text Processing, 530-541

Bensalem, I., Rosso, P., Chikhi, S. (2013). A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection. In: Forner, P., Müller, H., Paredes, R., Rosso, P., and Stein, B. (eds.) CLEF 2013, LNCS, vol. 8138. pp. 53–58. Springer, Heidelberg .

Boudelaa, S., and Marslen-Wilson, W. D. (2010). ARALEX: A lexical database for Modern Standard Arabic. Behavior Research Methods, 42, 481-487.

Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. LDC Catalog No.: LDC2002L49.

Doumi, N., Lehireche, A., Maurel, D. and Cherif, A-M. (2013). La conception d'un jeu de ressources libres pour le TAL arabe sous Unitex. In Proceedings of the the 6th international Colloquim TRADETAL, Oran, Algeria.

Dukes K. and Habash, H. (2010). Morphological Annotation of Quranic Arabic. In Proceedings of the Language Resources and Evaluation Conference (LREC). Valletta, Malta.

El-Haj, M., Koulali, R. (2013). KALIMAT a Multipurpose Arabic Corpus at the Second Workshop on Arabic Corpus Linguistics (WACL-2).

El-Haj, M., Kruschwitz, U. and Fox, C. (2010). Using Mechanical Turk to Create a Corpus of Arabic Summaries. In Proceedings of the The 7th International Language Resources and Evaluation Conference (LREC 2010), Valletta, Malta,. LREC.

El-Haj, M., Rayson, P. Using a Keyness Metric for Single and Multi Document Summarisation. In Proceedings of the MultiLing 2013 Workshop, held within the ACL 2013 Conference, Sofia, Bulgaria.

Elkateb S., Black, W., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., Fellbaum, C. (2006). Building a WordNet for Arabic. In Proceedings of the Fifth International Conference on Langauge Resources and Evaluation, Genoa, Italy.

Graja, M., Jaoua, M. and Hadrich Belguith, L. (2010) Lexical Study of A Spoken Dialogue Corpus in Tunisian Dialect. In Proceedings of the ACIT 2010: the International Arab Conference on Information Technology, Benghazi, Libya, 14–16 December 2010.

Habash, N. (2010). Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies.

Hassaïne, A. and Al Maadeed, S. (2012). ICFHR2012 Competition on Writer Identification – Challenge 2: Arabic Scripts. In In Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition, pp. 831-836, 2012.

Hassaïne, A., Al-Maadeed, S., Alja'am, J.M., Jaoua, A., Bouridane, A. (2011). The ICDAR2011 Arabic Writer Identification Contest. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pp.1470-1474, 18-21 Sept. 2011.

Khoja S., Garside, R. and Knowles, G. (2001). An Arabic Tagset for the Morphosyntactic Tagging of Arabic. In Corpus Linguistics 2001, Lancaster University, Lancaster, UK, March 2001, and to appear in a book entitled A Rainbow of Corpora: Corpus Linguistics and the Languages of the World, edited by Andrew Wilson, Paul Rayson, and Tony McEnery; Lincom-Europa, Munich.

Mesfar, S., Silberztein, M. (2008). Transducer Minimization and Information Compression for NooJ Dictionaries. In Proceedings of the FSMNLP 2008: 110-121.

Mezghani, S., Hammami, Hadrich Belguith, L., Ben Hamadou, A. (2009). Arabic Anaphora Resolution: Corpora Annotation with Coreferential links. In Proceedings of the International Arab Journal of Information Technology, vol. 6, No. 5, pp481-489, November 2009.

Mohit, B., Schneider, N., Bhowmick, R., Oflazer, K and Smith, N.-A. (2012). Recall-Oriented Learning of Named Entities in Arabic Wikipedia. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 162-173.

Mousser, J. (2010). A Large Coverage Verb Taxonomy for Arabic. 2010. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds.), LREC, European Language Resources Association. ISBN: 2-9517408-6-7.

Rafalovitch, A. and Dale R. (2009). United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In Proceedings of the MT Summit XII, pages 292-299, Ottawa, Canada.

Saad, M., Ashour, W. (2010). OSAC: Open Source Arabic Corpora. In Proceedings of the EEECS'10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, , pp. 118-123.

Saad, M., Langlois, D. and Smaïli, K. (2013). Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities. Procedia - Social and Behavioral Sciences, 95 (0): 40-47.

Schneider, N., Mohit, B., Oflazer, K. and Smith, N-A. (2012). Coarse Lexical Semantic Annotation with

Supersenses: An Arabic Case Study. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, South Korea, July 2012.

Steinberger, R., Pouliquen, B., Kabadjov, M., Belyaeva, J. and Van der Goot, E. (2011). JRC-Names: A freely available, highly multilingual named entity resource. In Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP). Hissar, Bulgaria, 12-14 September 2011.

Trigui, O., Hadrich Belguith, L. and Rosso, P. (2010). DefArabicQA, Arabic Definition Question Answering System. In Proceedings of the Workshop on Language Resources and Human Language Technologies for Semitic Languages, Language Resources and Evaluation Conference (LREC), May 17th 2010, Valletta, Malta.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., Houston, A. (2013). OntoNotes Release 5.0. LDC Catalog No.: LDC2013T19.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N.,Rozovskaya, A., Farra, N., Alkuhlani, S. and Oflazer, K. (2014). In Proceedings of the 9th Language Resources and Evaluation Conference (LREC), 26-31 May 2014, Reykjavik, Iceland.

Automatic Readability Prediction for Modern Standard Arabic

Jonathan Forsyth
Department of Linguistics and English Language
Brigham Young University, Provo, UT 84602 USA
jon4syth@gmail.com

Abstract

Research for automatic readability prediction of text has increased in the last decade and has shown that various machine learning (ML) methods can effectively address this problem. Many researchers have applied ML to readability prediction for English, while Modern Standard Arabic (MSA) has received little attention. Here I describe a system which leverages ML to automatically predict the readability of MSA. I gathered a corpus comprising 179 documents that were annotated with the Interagency Language Roundtable (ILR) levels. Then, I extracted lexical and discourse features from each document. Finally, I applied the Tilburg Memory-Based Learning (TiMBL) program to read these features and predict the ILR level of each document using 10-fold cross validation for both 3-way and 5-way classification tasks. I measured performance using the F-score. For 3-way and 5-way classifications my system achieved F-scores of 0.719 and 0.519 respectively. I discuss the implication of these results and the possibility of future development.

Keywords: Arabic, readability, machine learning

1. Introduction

In general, readability is the level of difficulty that a particular document presents to readers. Readers may be adults or children, and may be native or non-native speakers of the language in question. This compounds the difficulty for precisely defining readability, because readers and documents vary widely. Consequently, researchers in readability have developed many different ways of defining and measuring readability.

Formal second language education utilizes many reading materials. Textbooks alone do not provide enough material in order to learn to read fluently in the second language. The challenge, then, is to match other texts with learner's reading ability.

One solution is to have language instructors select for the students material which is appropriate to their reading ability. This requires the instructor to spend significant time analyzing documents to find ones at a target readability level. Also, given several students who have varied reading proficiencies, it is impractical for the instructor to choose material at multiple reading levels.

A more efficient solution would be to automate this process. The last decade has seen application of computerized methods to the readability problem for several languages including English, French, Spanish, and German. These methods show that automatic readability measurement can, in some cases, achieve human-like performance. Furthermore, readability automation could allow an independent second language learner to determine which documents they should read for the most effective learning experience.

Second language learners of Arabic do not yet have such tools available to them. Arabic includes a diverse collection of dialects across the Middle East and Northern Africa. MSA is the standard written form of Arabic and is heavily emphasized in second language education. My research focuses on automated readability prediction for MSA which is morphologically and syntactically complex and is often challenging for learners to acquire.

MSA newspaper writing provides a variety of vocabulary and can be useful for second language acquisition (Parkinson, 2006). However, the linguistic difficulty of many documents can render newspaper documents less effective in engaging the student in reading. Additionally, learners' competence in the subject area can be problematic. MSA newspaper concepts range from topics of political science and international relations to culture and business, among others. It is likely that with the automation of readability prediction, news sources that are accessible to students can be more easily discovered.

Many linguistic features have proven to be effective in predicting readability for various languages. Most of these have not been applied to MSA. Also, due to its complexity, features that are unique to MSA should be investigated. Some recent studies exist in this area, but more work is needed.

2. Overview

In Section 3 I review previous readability research beginning with English. I include research for other languages and discuss the recent application of ML to the readability problem. I also discuss various features. Then, in Section 4 I report research in MSA that relates to previous readability research. Next, in Section 5 I describe the resources and methods which I utilized to build my readability prediction system. In Section 6 I explain my procedures for evaluation. Finally, in Section 7 I present and discuss the performance results of my MSA readability system. I conclude in Section 8 with shortcomings of my research and prospects for further MSA readability research.

3. Review of the Literature

Within readability research, English has been the principal language of investigation, although scholars have researched readability in several other languages, including French (François and Watrin, 2011), German (Hancke et al., 2012), and even MSA (Al-Khalifa and Al-Ajlan, 2010). Research in non-English languages has shown that the factors affecting English readability can be useful for readability research in other languages. These factors have a long tradition in English readability research.

In the 1940s many readability researchers developed formulas to improve and formalize readability measurements for English; the common features in these formulas were measurements of the average length of sentences and words—known as traditional features (Flesch, 1948; Dale and Chall, 1948). Other common features were ratios of a document's words found in word lists of frequent or common vocabulary. These early methods only accounted for a few features because they required human counts.

In one of the first applications of ML to readability prediction, Si and Callan (2001) demonstrated that traditional formulas do not scale well to web documents. Traditional formulas need more data than they are afforded in many cases by new kinds of non-traditional documents such as web pages. Several studies have applied ML algorithms to readability prediction since this study as well as addressed the need for web search results to be automatically filtered for readability.

Vajjala and Meurers (2012), for example, studied the application of ML to readability prediction for texts from an online educational site—Weekly Reader.[1] They used several ratios of tokens, each with a particular part-of-speech (POS) as features of their model. Among their POS-ratio features, nouns were the most predictive of readability.

Discourse connectives, words and multi-word expressions that connect units of text larger than single words, are a recent feature in readability research. The English word 'however' is an example of a discourse connective that is often used to connect two sentence units. Pitler and Nenkova (2008) showed that discourse connective features perform well in readability scoring and readability ranking for English. They created an automated English readability assessment model that matched human ratings of various news articles. They used a large feature set and found document length and discourse relations between clausal arguments of explicit discourse connectives to be the most significant factors in training a model to match discrete human ratings.

4. Arabic Research

Discourse connectives are also an important linguistic trait of MSA writing. Al-Batal (1990) wrote that "MSA seems to have a connecting constraint that requires the writer to signal continuously to the reader, through the use of connectives, the type of link that exists between different parts of the document. This gives the connectives special importance as text-building elements and renders them essential for the reader's processing of text" (p. 256).

Alsaif and Markert (2011) developed the first model for automatic discourse connective identification for MSA. Their purpose was to disambiguate the authentic use of discourse forms as functional discourse connectives from the same forms that were not functional as discourse connectives. They based the features of their model on the Leeds Arabic Discourse Treebank (LADTB), a corpus of Arabic news that they annotated, (Alsaif and Markert, 2010), which is an excerpt of the Penn Arabic Tree Bank (Maamouri and Bies, 2004). In comparing their corpus with the PDTB they discovered that Arabic writers make use of explicit discourse connectives much more frequently in the newspaper genre than is found in the same genre for English. Therefore, discourse connectives may be very valuable in matching human ratings as Pitler and Nenkova show that they are in English. As part of their research they compiled a comprehensive discourse connective list available online.[2] Since discourse connectives proved useful as readability indicators in Pitler and Nenkova's study (2008), they may also be effective for Arabic readability.

Arabic readability research is in its very early stages; Shen et al. (2013) published a recent example. They use ML to create general readability classifiers for 4 languages: Arabic, English, Dari, and Pashto. They employ documents from the Defense Language Institute (DLI) Foreign Language Center as their corpus. These documents are annotated for readability according the Interagency Language Roundtable (ILR) system (Clark and Clifford, 1988). The ILR system consists of 11 proficiency levels for listening, speaking, reading, and writing which were developed by various U.S. government agencies. (See Table 1 adapted from the site of the ILR.[3]) The Arabic section of their corpus consists of 1,394 documents across 7 of the 11 ILR levels: 1, 1+, 2, 2+, 3, 3+ and 4—approximately 200 documents per class. They divide the Arabic section of their corpus into an 80/20 training/testing split.

Their feature set is small and includes traditional, language-independent features. The 2 categories of features they label are: 1) word usage and 2) shallow length features. The word usage category only includes weighted word frequencies from their training corpus. The shallow length category includes average sentence length in words, number of words per document, and average word length in characters. They normalize these three length features using a method to make the scores more comparable. They report results in terms of root mean squared error achieving 0.198 for Arabic among the 7 levels.

Level	Description
0	No Proficiency
0+	Memorized Proficiency
1	Elementary Proficiency
1+	Elementary Proficiency, Plus
2	Limited Working Proficiency
2+	Limited Working Proficiency, Plus
3	General Professional Proficiency
3+	General Professional Proficiency, Plus
4	Advanced Professional Proficiency
4+	Advanced Professional Proficiency, Plus
5	Functionally Native Proficiency

Table 1: *Interagency Language Roundtable (ILR) levels*

In another Arabic readability study, Al-Khalifa and Al-Ajlan (2010) create a document corpus from educational materials for elementary, intermediate, and sec-

[1] http://www.weeklyreader.com

[2] http://www.arabicdiscourse.net/annotation-tool/

[3] http://www.govtilr.org/Skills/ILRscale4.htm

ondary schools in Saudi Arabia. Their corpus has 150 documents—50 documents from each level—comprising 57,089 tokens. They process the documents of their corpus to generate 5 features: average sentence length, average word length in letters and syllables, term frequency (ratio of duplicated words), and a bigram language model. They build 3-way ML classifiers to classify the documents in this corpus across the 3 education levels. For evaluation they use an 80/20 training/testing split. They compare performance between two different sets of features—their entire feature set and a subset of the best three features: average sentence length, the bigram language model, and term frequency. They use F-scores to report performance against individual levels and accuracy for overall performance. I summarize their results in Table 2. Their results are much better than a baseline of classifying the documents at random—approximately 33% accuracy.

Level	All Features	Feature Subset
Easy	1.00	1.00
Medium	0.545	0.667
Difficult	0.615	0.667
Accuracy	72.22%	77.78%

Table 2: *F-scores and accuracy for classification using all features and the best-performing subset*

I've explained where readability research started and how it has advanced to using computational methods to automatically predict readability level. I also looked at discourse features which I believe will be useful in deriving readability levels automatically from MSA documents. In the following section I will discuss the tools I used for extracting useful linguistic information from MSA documents and for computing accurate predictions for readability levels of the same.

5. Resources and Methods

I downloaded a corpus in May 2013 from the online curriculum of the Defense Language Institute (DLI) Foreign Language Center.[4] The DLI corpus contains documents which are based on authentic MSA materials and are annotated with one of five ILR levels: 1, 1+, 2, 2+, and 3. The corpus has a total of 179 documents and 67,532 tokens excluding punctuation. The distribution of corpus documents across these levels is not equal as shown in Table 3.

Level	Train	Evaluation	All
1	16	4	20
1+	13	3	14
2	64	16	80
2+	32	8	40
3	20	5	25
Total	143	36	179

Table 3: *DLI corpus document levels and distributions*

I randomly partitioned the DLI corpus into two data sets: train and evaluation. The train set comprised 80% of the corpus with 52,873 tokens, and the evaluation set comprised the remaining 20% with 14,659 tokens. The number of documents in each set can also be seen in Table 3. This partitioning allowed for development and improvement of the classifier on the train set while preserving the integrity of the final results achieved with the evaluation set.

Previously, I referred to the use of features in readability research which are derived from different types of word lists. Modern frequency lists are more robust than early ones because they are typically based on multi-million word corpora. For my work I used 'A Frequency Dictionary of Arabic' (Buckwalter and Parkinson, 2011), a 5000-word frequency dictionary based on a 30-million word corpus.

Most of the features which I extracted from the corpus documents were based on the lemma, POS, rank, raw frequency, and range provided by the frequency dictionary. In order to identify and categorize words in a way which would allow me to compare them to the frequency dictionary, I required an Arabic morphological processing tool.

Such a tool can clarify MSA usage, which is often ambiguous in its written form, largely because diacritics are typically omitted. Diacritics represent short vowels, elongated consonants, and other distinguishing linguistic properties. MADA is a state-of-the-art program that accomplishes morphological analysis and disambiguation for MSA (Habash et al., 2009).[5] MADA provides full morphological disambiguation, POS tagging, English glosses and other useful information.

MADA's morphological analysis specifies, for each surface form, up to four proclitics, the lemma, and a possible enclitic. It includes 22 other features such as POS, full diacritization, and distinctive lexeme code. In addition to these 22 features, MADA outputs a user customized tokenization of the original input document. I obtained all of my features with the support of MADA's output. I included lexical and discourse features in my model. Lexical features are a strong and consistent feature in virtually all readability research and include measurements based on word length, word lists, and N-gram language models. I employed word lengths, a word list, and homographs in my model—the latter is a novel feature. I also used discourse connectives and word ratios. A full list of the features can be found in my previous work (Forsyth, 2014).

I used the comprehensive list of discourse connectives which Alsaif and Markert created (Alsaif, 2012). For simplicity, I employed the forms which they found to be used more than 50% of the time as authentic discourse connectives in their corpus. Coordinated connective pairs, such as the English 'if/then' construction are found in MSA also, but I excluded these from my subset of discourse connectives, because they are non-trivial to identify automatically.

Most of the lexical features I used relied on Buckwalter and Parkinson's frequency dictionary (2011). I normalized the dictionary entries to their equivalent MADA lexeme code in order to match them with MADA's output. After counting all frequent tokens, I divided this count by the total number of tokens in the document to obtain the frequency ratio

[4] http://gloss.dliflc.edu/Default.aspx

[5] I used the Linguistic Data Consortium's Standard Arabic Morphological Analyzer (SAMA) version 3.1 in connection with MADA.

feature.

The type-to-token ratio feature is common to several studies and I included it in my experiments. For English readability research the type-to-token ratio is calculated by counting the number of unique word types, and dividing them by the count of total tokens. I computed the type-to-token ratio by counting the MADA-produced lexeme code of each word, then dividing this count by the unique lexeme types. This decision excluded clitics and affixes from the type-to-token ratio.

I used the POS-based ratio features inspired by Vajjala and Meurers, but adjusted them to match MSA POS classes provided by MADA. As a reminder, Vajjala and Meurer's POS-based ratio features were the ratio of the individual POS type occurrences to the token count in the document. I found these to improve results when taken as a whole.

I demonstrated above that both modern and traditional readability methods have included average sentence length measures with positive results. I generated two measures of sentence length using both the average number of tokens and average number of morphemes per sentence. I used punctuation to delineate sentences, as the authors of the documents in this corpus were apparently consistent in their use of punctuation to mark sentence boundaries.

Another set of lexical features I included, which are novel for readability, are the homograph features. I identified homographs according to the lemmas in the frequency dictionary that had two or more entries for the same surface form. For example, the surface form /mA/ contains nine entries in the frequency dictionary, each with a different sense. This effectively limited the homographs to those that are frequent. The features I derived from these homographs were based on the number of entries in the dictionary and the number of instances in a given text.

All the features I included in my readability prediction system are summarized in Table 4. More than half of my features were frequency-based. I had 165 features altogether.

Feature Category	# of Features
POS-based Frequency Features	96
Type-To-Token POS Ratio Features	23
Token & Type Frequency Features	19
Discourse Connective Features	7
Homographic Features	5
Frequency-based Discourse Connective Features	4
Type-To-Token Features	4
Word Length Features	3
Sentence Length Features	2
Token Count Feature	1
Foreign Word Feature	1

Table 4: Features grouped by category and counts

I developed a Perl script that composes the features into an appropriate format for TiMBL—a feature vector. Each document's feature vector is inserted into a training or testing file for input. Then, the ML component extrapolates a reading level based on the comparison of each testing document to the training documents in memory.

I collected the features above for use in the TiMBL ML system which uses a strategy known as memory-based learning (Daelemans et al., 2010).[6] Memory-based learning relies on the theory that humans classify new information using analogy and proximity to the experiences they retain in memory. This study is novel in applying TiMBL to automated readability and, specifically, to MSA documents.

6. Evaluation Approach

In order to evaluate my automated readability system, I used 10-fold cross validation. This approach allows one to test the entire data set without bias and maintains a relatively large training sample. This is especially helpful for small corpora such as mine. To apply 10-fold cross validation I divided the documents of each level randomly into 10 partitions or folds, as is common in ML applied to readability. I configured TiMBL to leave one fold out as testing data (10%) and to include the remaining folds as training data (90%). This is repeated every 10 trials where each fold takes a turn as the test data. I also iterated this cross validation procedure 10 times with a random set of documents assigned to each fold each time. This provided a better sample across the documents than a single 10-fold cross validation iteration could. Finally, I averaged all results over all 10-fold cross validation iterations to report here. I automated this process with computer scripting and programming.

I tuned TiMBL classifiers using 10-fold cross validation on the train data set before applying them to the evaluation data set. Tuning involved running a trial and adjusting TiMBL settings and the included features based on the results to improve future results. I also tried splitting the train data in another 80/20 split for tuning purposes, but this was not as useful for tuning the classifiers because the testing partition's small size poorly represented the data set as a whole. I found that the results of repeated cross validation trials were more indicative of performance across samples. After tuning on the train set, I combined the train and evaluation sets and performed 3-way and 5-way 10-fold cross validations with 10 iterations each.

I used 162 of the 165 total features. Table 5 lists the features that I excluded. The range feature refers to the range measure provided in the frequency dictionary; it was too sparse since each document is very likely to have a lexical item with a range of 100—an item found in all subcorpora which comprise the base corpus of the frequency dictionary. The other two excluded features were both based on ranks of highly frequent lexical items which led to virtually no distinction in this measure across document levels.

Feature Name	Description
maxTypeFreqRange	max frequent word type range
minFreqPrepRank	min frequent preposition rank
minTypeFreqRank	min frequent word type rank

Table 5: *Excluded features for the final experiment*

Timbl accepts a wide range of settings. The settings I employed in all of the experiments below were the same. I

[6] I used TiMBL version 6.4.4.

chose the settings based on experimental tuning of the parameters.

For all 10-fold cross validation experiments, I added the counts for true positive, false positive, and false negative predictions across each iteration. Then, I calculated the precision, recall, and F-score for each class using these cumulative counts. Finally, I calculated the average F-score across the class F-Scores which I obtained. I report details of the final evaluation results in the following section.

7. Results

My preliminary evaluations provided a comparison to final evaluations. These preliminary evaluations treated the train set and included 3-fold, 5-fold, and 10-fold cross validation, as well as leave-one-out for both 3-way and 5-way classifiers. I only report the preliminary experiments for 3-way and 5-way 10-fold cross validation here as they showed the best performance. For the final evaluations I also ran 3-fold and 5-fold cross validations. In all experiments I used 162 features, leaving out 3 original features that proved ineffective due to their rarity. I did not test every possible combination of the 162 features, which would have been impractical even if I discounted the various TiMBL settings that could be set for each experiment.

In the preliminary evaluations my system had already 'seen' all the train set data, and I was able to improve the performance by adjusting the features and TiMBL settings incrementally. I found through this process that 10-fold cross validation produced the best results among all evaluations using the train set. Each of the preliminary results shown in Tables 6 and 7 represents the highest average F-scores I achieved with various features and TiMBL settings.

Level	Precision	Recall	F-Score
1 & 1+	0.840	0.777	0.807
2	0.724	0.810	0.765
2+ & 3	0.825	0.736	0.778
		Average F-Score	0.783

Table 6: *10-fold cross validation 3-way classifier—preliminary evaluation*

Level	Precision	Recall	F-Score
1	0.764	0.487	0.595
1+	0.602	0.427	0.500
2	0.558	0.760	0.644
2+	0.292	0.184	0.226
3	0.698	0.615	0.654
		Average F-Score	0.523

Table 7: *10-fold cross validation 5-way classifier—preliminary evaluation*

The results in Tables 6 and 7 show that the 3-way classifier performed the best, as expected because there are fewer levels to choose between. Results are similar for all levels. The 5-way classifier showed more variance between the individual levels. Notably, it performed very poorly on level 2+. The reason for this unimpressive performance may be that level 2+ documents were very similar to the neighboring levels. The overall F-score is barely above a baseline of choosing level 2 for each prediction which would achieve 44.7% accuracy—the F-score and accuracy approximate each other—as this is the percentage of level 2 documents in the train set.

Final results of 3-way and 5-way classifiers are shown in Tables 8 and 9 respectively. The 3-way classifier performed best with an F-score of 0.719. This result is nearly a 30% improvement over the baseline of 44.7% when choosing level 2 every time—the same as in preliminary experiments and the same here for a 5-way classifier. The individual level statistics for the 5-way classifier show that TiMBL predicted levels 1, 2, and 3 much more accurately than levels 1+ and 2+. The advantage seems to be partly that levels 1+ and 2+ each have two immediate neighboring levels to be distinguished from, whereas levels 1 and 3 must be distinguished from only one immediate neighbor. The better performance of level 2 (which also has two adjacent levels) is likely because of abundant training examples. Level 2 had the most documents available for training.

Level	Precision	Recall	F-Score
1 & 1+	0.818	0.582	0.680
2	0.663	0.795	0.723
2 & 2+	0.791	0.718	0.753
		Average F-Score	0.719

Table 8: *10-fold cross validation 3-way classifier*

Level	Precision	Recall	F-Score
1	0.813	0.610	0.697
1+	0.724	0.150	0.248
2	0.653	0.856	0.740
2+	0.460	0.290	0.355
3	0.503	0.624	0.557
		Average F-Score	0.519

Table 9: *10-fold cross validation 5-way classifier*

The difference in performance between the preliminary and final evaluations is such that the 3-way classifiers' F-scores decreased from 0.783 to 0.719. Likewise, the 5-way classifiers' F-score decreased from 0.523 to 0.519.

My automated readability prediction system achieved the best results with 3-way classifiers using all but 3 features which were found to be sparse in the train set. This is slightly lower than the results achieved by Al-Khalifa and Al-Ajlan (2010) in their study for with a 3-way classifier—77.78% accuracy which approximates the F-score. These two results, though, are not entirely comparable because they derive from different corpora. The DLI Corpus is written for adult readers of Arabic as a second language, while Al-Khalifa and Al-Ajlan used a grade-school corpus.

8. Conclusion

Readability has been an important research problem for nearly a century, though research into readability for MSA is in the very early stages. This work is an important contribution to the state-of-the-art for automated MSA readability prediction. It is the first study to employ TiMBL in readability prediction for any language and shows that this can be done to good effect in MSA.

A limitation of my study was that I used a very small corpus. Larger corpora are advantageous because they provide more training data for ML classifiers. The classifiers can be improved thereby since more training data is more representative of a whole data set. I anticipate that, as for other languages, more MSA corpora annotated for readability will be made available in the future.

My system does not include syntactic features. MSA has very rich syntax that reflects a wide range of complexity. I excluded syntactic features because of the difficulty in applying syntactic parsing to my data set. Syntactic features have shown positive contributions in previous readability research. Ongoing work on Arabic parser development will render them more robust to data like mine.

A very helpful application for future work would be a graphical program for inputting MSA documents, to produce a readability score. This could be made widely available to students and teachers in a web-based context. Similarly, web search engines could build upon this research to return results according to specific readability levels.

In conclusion, I developed the first TiMBL-based system for distinguishing MSA documents annotated with ILR readability levels. I used some novel features some of which were derived from a frequency dictionary. This is the only MSA readability prediction system that uses a frequency dictionary. Using standard ML evaluation techniques I was able to produce positive results, even with a smaller corpus than used in similar studies. My study has shown that the combination of several lexical, discourse, and traditional features are effective indicators of MSA readability and that further research to improve MSA readability is worthwhile.

9. References

Al-Batal, M. (1990). Connectives as cohesive elements in a modern expository Arabic text. In Eid, M. and McCarthy, J., editors, *Perspectives on Arabic Linguistics II*. John Benjamins, Amsterdam/Philadelphia.

Al-Khalifa, H. S. and Al-Ajlan, A. A. (2010). Automatic readability measurements of the Arabic text: An exploratory study. *The Arabian Journal for Science and Engineering*, 35(2C):103–124.

Alsaif, A. and Markert, K. (2010). The leeds Arabic discourse treebank: Annotating discourse connectives for Arabic. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC)*, pages 2046–2053. European Language Resources Association.

Alsaif, A. and Markert, K. (2011). Modeling discourse relations for Arabic. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 736–747. Association for Computational Linguistics.

Alsaif, A. (2012). *Human and Automatic Annotation of Discourse Relations for Arabic*. Ph.D. thesis, University of Leeds, Leeds, UK.

Buckwalter, T. and Parkinson, D. (2011). *A Frequency Dictionary of Arabic: Core Vocabulary for Learners*. Routledge Frequency Dictionaries.

Clark, J. L. and Clifford, R. T. (1988). The FSI/ILR/ACTFL proficiency scales and testing techniques: Development, current status, and needed research. *Studies in Second Language Acquisition*, 10(2):129–147.

Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. (2010). TiMBL: Tilburg Memory Based Learner, version 6.3, reference guide. Technical Report version 6.3, ILK Research Group Technical Report.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Forsyth, J. (2014). Automatic readability prediction for Modern Standard Arabic. Master's thesis, Brigham Young University, Provo, UT.

François, T. and Watrin, P. (2011). On the contribution of MWE-based features to a readability formula for French as a foreign language. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 441–447. Association for Computational Linguistics.

Habash, N., Rambow, O., and Roth, R. (2009). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, pages 102–109. The MEDAR Consortium.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING): Technical Papers*, pages 1063–1080. Association for Computational Linguistics.

Maamouri, M. and Bies, A. (2004). Developing an Arabic treebank: Methods, guidelines, procedures, and tools. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (COLING)*. Association for Computational Linguistics.

Parkinson, D. (2006). *Using Arabic synonyms*. Cambridge University Press.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Linguistics*, pages 186–195. Association for Computational Linguistics.

Shen, W., Williams, J., Marius, T., and Salesky, E. (2013). A language-independent approach to automatic text difficulty assessment for second-language learners. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 30–38. Association for Computational Linguistics.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*, pages 574–576. Association for Computing Machinery.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of The 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 163–173. Association for Computational Linguistics.

Subjectivity and Sentiment Analysis of Arabic Twitter Feeds with Limited Resources

Eshrag Refaee and Verena Rieser
Interaction Lab, Heriot-Watt University,
EH144AS Edinburgh, Untied Kingdom.
eaar1@hw.ac.uk, v.t.rieser@hw.ac.uk

Abstract

This paper addresses the task of automatic Subjectivity and Sentiment Analysis (SSA) for Arabic tweets. This is a challenging task because, first, there are no freely available annotated corpora available for this task, and second, most natural language processing (NLP) tools for Arabic are developed for Modern Standard Arabic (MSA) only and fail to capture the wide range of dialects used in Arabic micro-blogs. In the following paper we show that, despite these challenges, we are able to learn a SSA classifier from limited amounts of manually annotated data, which reaches performance levels of up to 87.7% accuracy using cross-validation. However, an evaluation on a independent test set shows that these static models do not transfer well to new data sets, collected at a later point in time. An error analysis confirms that this drop in performance is due to topic-shifts in the twitter stream. Our next step is to extend our current models to perform semi-supervised online learning in order to continuously adapt to the dynamic nature of online media.

Keywords: Subjectivity and Sentiment Analysis, Arabic, Twitter, Learning from small data sets

1. Introduction

Compared to other languages, such as English, research on Arabic text for Subjectivity and Sentiment Analysis (SSA) is sparse. One possible reason is that Arabic is still an under-resourced language in the Natural Language Processing (NLP) community, mainly because of the complex morphological, structural, and grammatical nature of Arabic (Habash, 2010). Furthermore, annotated corpora for Arabic SSA are not freely available. While there is a growing interest within the NLP community to build Arabic corpora by harvesting the web, e.g. (Al-Sabbagh and Girju, 2012; Abdul-Mageed and Diab, 2012; Zaidan and Callison-Burch, 2013; Mourad and Darwish, 2013), these resources have not been publicly released yet. We therefore utilise a newly collected corpus of annotated twitter feeds, which is released via the ELRA repository (Refaee and Rieser, 2014a). We develop a automatic SSA classifier on this data set and explore the effectiveness of different feature-sets, as applied by previous work, e.g. (Abdul-Mageed et al., 2011). We show that features automatically extracted with freely available Arabic NLP tools (developed for standard Arabic only) can achieve performance levels that are positively related to classification results.

1.1. Background

Arabic is the language of an aggregate population of over 422 million people, first language of the 22 member countries of the Arabic League and official language in three others (UNESCO, 2013). Arabic can be classified with respect to its morphology, syntax, and lexical combinations into three different categories: classic Arabic (CA), modern standard Arabic (MSA), and dialectal Arabic (DA) (Habash, 2010). Users on social networks typically use the dialects varieties/ Arabs' native tongue such as Egyptian Arabic and Gulf Arabic (Al-Sabbagh and Girju, 2012). Dealing with DA creates additional challenges for natural language processing (NLP); being mainly spoken, dialects lack standardisation, and are written in free-text (Zaidan and Callison-Burch, 2013).[1]

This problem is even more pronounced when moving to the micro-blog domain, such as twitter. People posting text on social networks tend to use informal writing style, for example by introducing their own abbreviations, as in example (1), or using spelling variations. Furthermore, bi/multilingual users tend to use a mixture of languages, as in example (2) taken from our corpus (see Section 2.). In addition, tweets may also convey sarcasm, mixed and/or unclear polarity content.

(1) *I'll be right back*
برب

(2) *football match* in English, spelled using the Arabic alphabet.
مَاتش كورة

In contrast to grammar- or lexicon-based approaches to SSA, machine learning techniques are in general robust to such variety, but require annotated corpora, which are sparse for Arabic SSA. We therefore experiment with learning classifiers on a newly collected small corpus of over 3k of annotated twitter feeds and we automatically extract a variety of features using freely available Arabic NLP tools. However, most freely available Arabic NLP tools to date, are developed for MSA only. In the following, we investigate whether their performance is sufficient for providing informative features for SSA classification. We find that amongst our best preforming features for SSA classification are token-based n-grams, confirming previous results, e.g. (Wilson et al., 2009; Abdul-Mageed et al., 2011).

1.2. Related Work

Previous work on SSA has used manually annotated gold-standard data sets to analyse which feature sets and models perform best for this task, e.g. (Wiebe et al., 1999);

[1] It is important to mention current efforts by (Habash et al., 2012) to establish a conventional orthography for dialectal Arabic (CODA) to define a standard orthography of DA. This will, in future work, facilitate the development of NLP tools for DA.

Type	Feature-sets
Morphological	Diacritic, Aspect, Gender, Mood, Person, Part_of_speech, State, Voice, has_morphological_analysis .
Syntactic	n-grams of words and POS, lemmas, including Bag_of_Words (BOW), Bag_of_lemmas.
Semantic	Has_positive_lexicon, Has_negative_lexicon, Has_neutral_lexicon, Has_negator
Stylistic	Has_positive_emoticon, Has_negative_emoticon.

Table 1: Annotated Feature-sets

(Wilson et al., 2009). Most of this work is in English, but there have been first attempts to apply similar techniques to Arabic (MSA), e.g. (Abdul-Mageed et al., 2011). So far, only few studies have investigated Arabic social media (Abbasi et al., 2008; Abdul-Mageed et al., 2012; Mourad and Darwish, 2013). Our work is closely related to (Abdul-Mageed et al., 2012; Mourad and Darwish, 2013): We evaluate whether features used for MSA, e.g. by (Abdul-Mageed et al., 2011), can be transferred to Arabic Tweets written in DA. In addition, we test our models on a held-out test set, collected at a later point in time to explore the performance of our models for a dynamic medium, such as twitter. In contrast, (Mourad and Darwish, 2013) only use cross-validation to evaluate there classifiers. (Abdul-Mageed et al., 2012) use a held-out test set, which is a subset of the same data set used for training, and thus, doesn't represent a real-world scenario, where trained models are used to classify a stream of twitter feeds over a period of time.

2. Arabic Twitter SSA Corpora

We describe a newly collected corpus of over 3k of twitter feeds manually annotated gold-standard SSA labels. We use the Twitter Search API for corpus collection, which allows harvesting a stream of real-time tweets by querying their content. The extracted data is cleaned in a pre-processing step, e.g. removing duplicates, normalising non-Arabic, digits, user names, and links.

We harvest two data sets at two different time steps, which we label manually. We use the development set to train and cross-validate classification models. The test set is used as independent held-out set for evaluation. The tweets were collected by querying the Twitter API for a number of entities i.e. *Egypt, Bashar, Valentine, and tourism* . The query terms were then replaced by place-holders to avoid bias.

Development data: This data set contains 3,309 multi-dialectal Arabic tweets randomly retrieved over the period from January 25th to March 5th 2013.

Test data: We also manually labelled a subset of 963 tweets, which were collected between 6th to 15th of November 2013.

Two native speakers recruited to manually annotate the collected data for subjectivity, i.e. subjective/polar versus objective tweets, and sentiment, where we define sentiment as a positive or negative emotion, opinion, or attitude, following (Wilson et al., 2009). Our gold-standard annotations reached an inter-annotator agreement of weighted Kappa $\kappa = 0.76$, which indicates reliable annotations (Carletta, 1996). To avoid duplicated data instances, re-tweeted tweets were removed from the dataset. Table 3 gives examples of annotated tweets. Table 2 shows the distribution of labels in the two data sets.

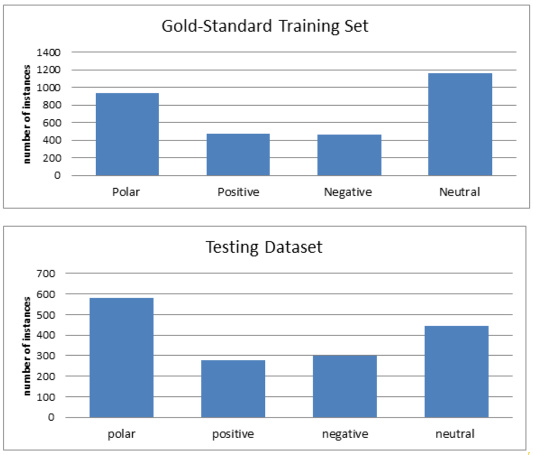


Table 2: Sentiment label distribution of the gold-standard manually annotated training and testing data sets.

Please note, that the development set described in this work cannot be released due to restrictions posed by Twitter Inc. However, a newly collected and annotated Arabic Twitter dataset that follows Twitter's regulations for publication will be released via ELRA repository (Refaee and Rieser, 2014a).

`postitive`	السيَاحه فِي اليمن جَمَال لَا يصدق	*Tourism in Yemen, unbelievable beauty*
`negative`	كميه الحَالَات النفسيه فِي شبَاب هذَا الجيل مخيفه جدًا و اعتقد انهَا اهم اسبَاب الطلَاق	*Psychological conditions in the young generation are very scary and I think it's one of the most important reasons of divorce*
`neutral`	ميركل تدعو اوكرَانيَا لتشكيل حكومه جديده	*Merkel calls for Ukraine to form a new government*

Table 3: Examples of annotated tweets

2.1. Features

We annotate the corpus with a rich set of linguistically motivated features, see Table 1, where a subset has been showing an increase in the performance of sentiment analysis on MSA news-wire texts (Abdul-Mageed et al., 2011). We employ morphological features, simple syntactic features,

such as n-grams, and semantic features from a freely available polarity lexicon (Abdul-Mageed et al., 2011).
For a more detailed descriptions of features please see (Refaee and Rieser, 2014a).

Syntactic Features/ Word Tokens: We experiment with lexical representations of 1st, 2nd, and 3rd order of word-based n-grams.

Morphological Features Considering the morphological rich nature of Arabic, we annotate the following features: aspect, gender, mood (e.g. indicative), number, person, and voice (e.g. active). We utilise a state-of-art automatic morphological analyser for Arabic text to obtain these features. In particular, we incorporate the current version of MADA+TOKAN (v 3.2) developed by researchers at Colombia University (Habash and Rambow, 2005; Nizar Habash and Roth, 2009) which performs tokenization, diacrization, morphological disambiguation, Part-of-Speech (POS) tagging, stemming and lemmatisation for Arabic. It is important to mention that MADA is developed for Modern Standard Arabic (MSA) only. Tweets, in contrast, contain dialectal and/or misspelled words where the analyser is incapable of generating morphological interpretations. We therefore include a feature `has_morphological_analysis`.

Semantic Features: This feature set includes a number of binary features that check the presence of sentiment-bearing words of a polarity lexicon in each given tweet. To obtain this set of features, we exploit an existing manually annotated subjectivity lexicon, namely ArabSenti (Abdul-Mageed et al., 2011). In addition, we make use of a publicly available English subjectivity lexicon, MPQA (Wilson et al., 2009), which we automatically translate using Google Translate, following a similar technique to (Mourad and Darwish, 2013). The translated lexicon is manually corrected by removing translations with neutral or no clear sentiment indicator. For instance, *the day of judgement* is assigned with a negative label while its Arabic translation is neutral considering the context-independent polarity. This results in 2,627 translated instances after correction. We then construct a third dialectal lexicon of 484 words that we extracted from an independent Twitter development set and manually annotated for sentiment. All lexicons were merged into a combined lexicon of 4,422 annotated sentiment words and phrases (duplicates removed).

Stylistic Features: This feature-set includes two binary features that check the presence of positive/negative emoticons.

3. Experimental Setup

For classification, we experiment with two alternative problem formulations: Related work has treated subjectivity and sentiment analysis as two-stage binary classification process, where the first level distinguishes subjective and objective statements, and the second level then further distinguishes subjectivity into: subjective-positive / subjective-negative, e.g. (Wiebe et al., 1999; Abdul-Mageed et al., 2011; Abdul-Mageed et al., 2012). Alternatively, the classification can be carried out at as single-level classification (positive, negative, neutral), e.g. (Farra et al., 2010). We experiment with both options by collapsing the positive and negative labels of the 1-level into the polar label.
We experiment with a number of machine learning methods and we report the results of the best performing scheme, namely Support Vector Machines (SVMs), where we use the implementation provided by the WEKA data mining package version 3.7.9 (Witten and Frank, 2005). WEKA is a collection of machine learning algorithms for data mining tasks (Group, 2013). We compare our results against a majority baseline (i.e. the ZeroR classifier in WEKA). We report the results with two metrics: weighted f-score[2] and accuracy. We use paired t-tests to establish significant differences ($p < .05$).
For each experiment, we examine the effectiveness of our features for sentence-level sentiment classification. We start by using lemma-based uni-grams ("Bag-of-Words") as the basic performance, and then compare the increase in classification accuracy when adding additional features one-by-one. In future work, we will also experiment with more principled feature sub-set selection methods, following (Rieser and Lemon, 2006).
We also run a two-stage evaluation: For the first evaluation phase, we adopt 10-fold cross-validation on the development data. For the second phase of evaluation, we collect and annotate a new corpus to act as held-out set.

4. Classification Experiments

In the following, we first experiment with developing and testing our models using 10-fold cross-validation on the gold-standard data set, where we compare single-level vs. multiple level classification, see Sections 4.1. and 4.2. Most previous work stops here. In this work, however, we evaluate our best models on a independent held-out set, see Section 4.4. We show that, despite very promising cross-validation results (out-performing previous work on similar tasks), our models do not generalise well to data sets collected at a later point in time.

4.1. Hierarchical binary SSA classification

Polar vs. Neutral: As part of the hierarchical binary classification model, we first experiment with identifying neutral, i.e. objective, versus polar, i.e. subjective, statements. Results by (Wilson et al., 2009) indicate that the ability to recognise neutral classes in the first place, can greatly improve the performance for distinguishing between positive and negative utterances later on. Table 1 shows the feature-sets exploited in this set of experiments.
Table 4 summarises the results using different feature sets in combination with SVM models. All the classifiers significantly outperform the majority baseline. Adding additional feature sets did not have a significant impact on the performance achieved with bag-of-words (BOW) only, but shows a positive trend. The classifier reaches its highest performance of 86.93% accuracy using all features, which is a 0.09% absolute improvement over using BOW only. Adding morphological and semantic features slightly (but not significantly) hurts the performance.

[2] The weighted f-measure is the sum of all f-scores attained during cross-validation , each weighted according to the number of instances with that particular class label.

	Baseline		SVM	
Feature-sets	F	Acc.	F	Acc.
BOW	0.44	59.21	0.86	86.84
BOW + morph n-gram	0.44	59.21	0.86	86.75
BOW + morph n-gram + semantic	0.44	59.21	0.86	86.71
BOW + morph n-gram + semantic + stylistic	0.44	59.21	0.86	86.93

Table 4: 2-level classification: polar vs. neutral

	Baseline		SVM	
Feature-sets	F	Acc.	F	Acc.
BOW	0.67	50.22	0.88	87.74
BOW + morph n-gram	0.67	50.22	0.69	66.88
BOW + morph n-gram + semantic	0.67	50.22	0.69	65.81
BOW + morph n-gram + semantic + stylistic	0.67	50.22	0.87	87.10

Table 5: 2-level classification: positive vs. negative

Positive vs. Negative: As a second step in the hierarchical binary classification model, we distinguish between positive and negative sentiment. Table 5 summarises the results. Again, the bag-of-words significantly outperforms the majority baseline. Surprisingly, adding morphological significantly hurts the overall performance (20.86% absolute performance drop in accuracy). This confirms findings by (Abdul-Mageed et al., 2012), which observe that morphological information is more effective for subjectivity than for sentiment classification.

4.2. Single level multi-class classification

We now experiment with single level multi-class classification for SSA.

Positive vs. Negative vs. Neutral: The results in Table 6 show a negligible (1.62%) lower performance (in terms of accuracy) to the binary classification results, which we calculated by taking the average of both classification results for subjectivity and sentiment for the combination of all feature-sets.

Again, all classifiers outperformed the baseline. Here, adding morphological features has a significant positive effect on accuracy.

4.3. Summary of Cross-Validation Results

In summary, the cross-validation experiments revealed:

- Support Vector Machines in general outperformed the baseline for this task, confirming similar results by (Wilson et al., 2009; Abdul-Mageed et al., 2011). SVMs are known to perform well with large feature sets, which is especially relevant when using word-based features.
- The hierarchical classification model is only marginally better than the three-way classification model, which is surprising since multi-class classification is likely to be more difficult task than binary classification.
- Amongst our best performing feature sets are lemma-based uni-grams(“bag-of-words”, BOW). This confirms findings by (Wilson et al., 2009; Abdul-Mageed et al., 2011).
- Morphological features decreases the performance for sentiment analysis. We hypothesise that this might be due to the noise introduced by the morphological analyser MADA (Nizar Habash and Roth, 2009), which is designed for MSA only. Polar tweets, in contrast, are likely to contain dialectal words and expressions. In future work, we will exploit forthcoming tools designed specifically for Arabic Social Networks, e.g. (Al-Sabbagh and Girju, 2012).
- Adding the semantic and stylistic features has shown no significant impact to the performance. For the semantic features, mainly being lexicon-based, this might be a result of insufficient coverage of DA in the employed lexicon.

	Baseline		SVM	
Feature-sets	F	Acc.	F	Acc.
BOW	0.44	59.21	0.68	69.74
BOW + morph n-gram	0.44	59.21	0.76	76.67
BOW + morph n-gram + semantic	0.44	59.21	0.76	76.84
BOW + morph n-gram + semantic + stylistic	0.44	59.21	0.85	85.39

Table 6: Single-level classification: positive vs. negative vs. neutral

Whilst not directly comparable to our results, it still worth mentioning results reported by previous work on SSA twitter classification: (Abdul-Mageed et al., 2012) achieved an accuracy of 72.52% for subjectivity classification on Arabic tweets, and 65.87% on distinguishing sentiment, evaluating on 10% of the corpus as test data. In comparison, our best performing models using cross-validation achieve 86.93% for subjectivity and 87.10% for sentiment classification.

4.4. Results on Independent Test Set

We now test our best performing models (as identified using planned t-tests on the cross-validation results) on an independent held-out set, which was collected at a later point in time (see Section 2.). The purpose of this experiment is to evaluate the ability of our models to perform SSA classification for a time-changing platform like twitter. The results are summarised in Table 7. We can observe a significant performance drop of 38.88% on average between the cross-validation results and the results on independent test set. While it is common to observe a drop in performance on an independent test set, related work only reports of a drop of 4.64% on average on a similar task (Abdul-Mageed et al., 2011). In contrast to (Abdul-Mageed et al., 2011) our test

Data set	**Feature-set**	10-fold Cross-Valid.				Independent Test Set			
		Baseline		SVM		Baseline		SVM	
		F	Acc.	F	Acc.	F	Acc.	F	Acc.
Polar vs. Neutral	BOW + morph (1)-grams + semantic features + stylistic features	0.44	59.21	0.86	86.93	0.69	53.68	0.43	46.62
Positive vs. Negative	BOW	0.67	50.22	0.88	87.74	0.69	52.75	0.41	49.65
Positive vs. Negative vs. Neutral	BOW + morph (1+2)-grams + semantic features + stylistic features	0.44	59.21	0.85	85.39	0.63	46.31	0.28	28.24

Table 7: Selected evaluation results for the best performing models.

	Development Set (Spring'13)			**Test Set (Sept'13)**		
ID	**Arabic**	**English**	χ^2	**Arabic**	**English**	χ^2
1	الخير	well-being	10.0221	اجمل	more beautiful	7.061
2	الشعب	nation	7.114	احسن	better	5.8727
3	اجمل	more beautiful	6.9927	آه	(sigh)	5.236
4	وزير	minster	5.5587	سعَادة	happiness	4.689
5	مَاهر	skilful	5.0705	شهد	Honey	4.689
6	اهلك	annihilate/destroyed	5.0705	قلوب	Hearts	4.689
7	الدين	religion	5.0705	الخير	well-being	4.689
8	مَبروك	Congratulations	4.984	آسف	sorry	4.3405
9	بخير	fine	4.984	نحبك	love you	3.5099
10	حرَام	ill-gotten	4.052	اصحَاب	friends	3.5099

Table 8: The most predictive word uni-grams in the two data sets as evaluated by Chi-Squared.

set was collected about 8 months after collecting our development set. We hypothesise that this performance drop is caused by time dependent topic-shifts and the prominent role of n-gram lemma features in our models. Since twitter experiences topic-shifts over time, the vocabulary, especially the content words, are likely to change as well (Bifet and Frank, 2010; Go et al., 2009).

Error Analysis: We conduct a detailed error analysis in order to confirm this hypothesis. We investigate topic-shifts, by comparing the predictive power of word uni-grams in the different data sets as measured by Chi squared (χ^2) attribute evaluation, see Table 8. First, note that there is only 1/5 overlap between the predictive / discriminative words in the training and test sets. The word frequency distribution also differs amongst the two data sets: the overall overlap of unique tokens in only 12.21%.

Some of these observations might be an artefact of corpus size, since the development set about 2.4 times the size of test set. However, there is certainly a trend which confirms that topics shift over time and that our models trained on a small size manually annotated corpus are not general enough to support this shift.

5. Conclusion and Future Work

We address the task of automatic Subjectivity and Sentiment Analysis (SSA) for Arabic tweets. First, we collect and manually label an Arabic twitter corpus, which we automatically annotate with a rich set of features. We find that models trained on a small gold-standard data-set achieve considerably high performance levels of up to 87% accuracy using cross-validation. However, the performance significantly degrades when evaluating the models on a test set collected at a later point in time.

Confirming findings by previous work, e.g. (Wilson et al., 2009; Abdul-Mageed et al., 2011), lemma-based unigrams ("Bag-of-Words") are amongst our best performing features. However, this proves to be problematic when applying the learned models over time, since the topics, and thus the predictive lexical features, will have changed. We show this topic shift by analysing the most predictive unigram features in both data sets, supporting previous claims by (Bifet and Frank, 2010; Go et al., 2009). One possible solution is to annotate a larger data set. For example, (Abdul-Mageed et al., 2012) use a total of 11,000 gold-labelled social media instances. This amount of data is not only expensive to collect and annotate, but in addition, the learned models are not transferrable to new domains: Results by (Abdul-Mageed et al., 2012) suggest that individualised solutions are needed for each domain and task. As such, manual annotation of new data sets for each domain and task are not a feasible solution.

In current work (Refaee and Rieser, 2014b) we experiment with a semi-supervised technique called "distant supervision" where we use emoticons as (noisy) automatic labels for learning, following (Read, 2005).

Acknowledgements

The first author would like to thank the Saudi Arabian government for supporting her with a PhD scholarship.

6. References

Abbasi, A., Chen, H., and salem, A. (2008). Sentiment analysis in muliple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(1-34).

Abdul-Mageed, M. and Diab, M. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Abdul-Mageed, M., Diab, M. T., and Korayem, M. (2011). Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 587–591, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abdul-Mageed, M., Kübler, S., and Diab, M. (2012). Samar: A system for subjectivity and sentiment analysis of arabic social media. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 19–28. Association for Computational Linguistics.

Al-Sabbagh, R. and Girju, R. (2012). Yadac: Yet another dialectal arabic corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Bifet, A. and Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Farra, N., Challita, E., Assi, R. A., and Hajj, H. (2010). Sentence-level and document-level sentiment mining for arabic texts. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1114–1119. IEEE.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Group, M. L. (2013). WEKA3: Data Mining Software in Java. `http://www.cs.waikato.ac.nz/ml/weka/`.

Habash, N. and Rambow, O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Habash, N., Diab, M. T., and Rambow, O. (2012). Conventional orthography for dialectal arabic. In *LREC*, pages 711–718.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Mourad, A. and Darwish, K. (2013). Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. *WASSA 2013*, page 55.

Nizar Habash, O. R. and Roth, R. (2009). MADA+TOKAN: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In Choukri, K. and Maegaard, B., editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April. The MEDAR Consortium.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics.

Refaee, E. and Rieser, V. (2014a). An arabic twitter corpus for subjectivity and sentiment analysis. In *9th International Conference on Language Resources and Evaluation (LREC'14)*.

Refaee, E. and Rieser, V. (2014b). Can we read from a smiley face? emoticon-based distant supervision for subjectivity and sentiment analysis of arabic twitter feeds. In *9th International Conference on Language Resources and Evaluation (LREC'14)*.

Rieser, V. and Lemon, O. (2006). Using machine learning to explore human multimodal clarification strategies. In *International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*.

UNESCO. (2013). World Arabic Language Day. `http://www.unesco.org/new/en/unesco/events/prizes-and-celebrations/celebrations/international-days/world-arabic-language-day-2013/`.

Wiebe, J. M., Bruce, R. F., and O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Zaidan, O. F. and Callison-Burch, C. (2013). Arabic dialect identification. *Computational Linguistics*.

Designing, Building, and Analyzing an Arabic Speech Emotional Corpus

Ali Meftah[1], Yousef Alotaibi[2], Sid-Ahmed Selouani[3]

[1,2]College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[3]Université de Moncton, 218 bvd. J.-D.-Gauthier, Shippagan, E8S 1P6, Canada
{ameftah, yaalotaibi}@ksu.edu.sa, selouani@umcs.ca

Abstract

In this paper we describe a new emotional speech corpus recorded for Modern Standard Arabic (MSA). The newly designed corpus contains five target emotions, namely neutral, sadness, happy, surprised, and questioning, and it consists of 16 selected sentences that were read by 20 male and female native Arabic speakers. A human perceptual test was then applied to the recorded corpus. The test was performed by nine additional individuals . The results of the perceptual test verified the correctness of the intended emotions at a rate of 82.13%. Specifically, the most accurately identified emotion was "questioning," while the least identified emotion was "happy." Subsequent analyses of the results revealed that content sentences play an important role in influencing speakers with respect to controlling the intended emotion. This corpus is the first MSA corpus to be built using multiple variables, including gender, language content, and other speaker demographic data.

Keywords: Emotion, Arabic, Recording

1. Introduction

Recently, increased attention has been directed at the study of the emotional content of speech signals (El Ayadi, Kamel, & Karray, 2011). Understanding the emotions present in speech and synthesizing desired emotions in speech according to the intended message are the basic goals of emotional speech processing (Koolagudi & Rao, 2012). Different types of corpora have been used for the study of emotions in speech: corpora of spontaneous speech, corpora of acted speech, and corpora of elicited speech. Spontaneous speech corpora are very difficult to obtain, while acted speech corpora consist of texts read by a professional actor. Elicited speech corpora are created by putting a speaker into a situation meant to evoke a specific emotion (Navas et al., 2004). This paper presents a new Modern Standard Arabic (MSA) emotional speech corpus called the King Saud University KACST Text-To-Speech Database (KSUKTD) that we have designed, recorded, and evaluated.

1.1 Related Work

Burkhardt et al. (Burkhardt et al., 2005) recorded an emotional database in which 10 German speakers (five male and five female) simulated the seven emotions of *neutral*, *anger*, *fear*, *joy*, *sadness*, *disgust*, and *boredom*. They used 10 German utterances (five short and five longer sentences). The database was evaluated using a perceptual test. Kostoulas et al. (Kostoulas, Ganchev, Mporas, & Fakotakis, 2008) reported the design of a real-world emotional speech database for Modern Greek in which 43 speakers (23 male and 20 female) performed the six emotions of *delighted*, *pleased*, *neutral*, *confused*, *angry*, and *hot angry*. Navas et al. (Navas et al., 2004) described the designing and recording of an emotional speech database for standard Basque. The designed database contains six basic emotions performed by a professional dubbing actress. The recorded database is a total of 1 hour and 35 minutes in length. Saratxaga and Navas (2006) also described the designing and recording of an emotional speech database for standard Basque. The database was designed with the twofold purpose of being used for corpus-based synthesis and for allowing the study of prosodic models of emotions. The database consists of approximately 1.5 hours per emotion, comprising 10.5 hours of recordings per speaker, for a total of more than 20 hours. Bhutekar and Chandak (2012) described the factors used in designing and recording large speech databases for applications requiring speech synthesis. They focused on the factors affecting the design of recording prompts, speaker selection procedure, the recording setup, and the quality control of the resulting database. Engberg and Hansen (1997) created a database in which they recorded and analyzed Danish emotional speech. Four actors, two male and two female, recorded five emotions: neutral, surprise, happiness, sadness, and anger. The database comprised approximately 30 minutes of speech. A listening test with 20 listeners was conducted to test the database; the emotions were identified correctly in 67.3% of the cases, with a 95% [66.0–68.6] confidence interval. Slobodan et al. (2004) presented their results of the design, processing, and evaluation of a Serbian emotional speech database. Six actors, three male and three female, were used for the recording of five emotions: neutral, anger, happiness, sadness, and fear. The database consists of 32 isolated words, 30 short semantically neutral sentences, 30 long semantically neutral sentences, and one passage of 79 words in length. The authors claimed that the listening test showed the correct identification of emotions in 95% of the cases. Hozjan et al. (2002) reported the design of an emotional speech database for Slovenian, English, Spanish, and French for the general study of emotional speech. Six emotions were recorded: anger, sadness, joy, fear, disgust, and surprise. Neutral emotional styles were also recorded. Two actors, one male and one female, were recorded for all languages except English, for which two male speakers and one female speaker were used. For each language, 175–190 sentences were used. Staroniewicz & Majewski (2009) published a state-of-the-field review on emotional speech databases while also designing their own Polish database in which they recorded the six simulated emotional states of anger, sadness, happiness, fear, disgust, surprise, and neutral by speakers from three groups: professional actors, amateur actors, and amateurs. The authors also describe many general principles that are useful for determining naturalness, selecting emotions for testing, selecting speakers, selecting texts, and validating various procedures.

1.2 Objective and Motivation

Our goal is to design and create an Arabic emotional speech corpus. This corpus is then evaluated through a rigorous perceptual test. To do that a pool of human listeners are selected in order to perform emotion classification of the recorded speech through listening sessions. The results will allow us to assess the selected sentences, speakers and emotions and to make recommendations for a second phase of the emotional corpus design. The ultimate goal is to provide a reliable linguistic resource that will be very useful for the research in the field of emotional Arabic speech.

2. Corpus Design

The Arabic language in general suffers from a lack of speech corpora, and in particular emotional speech corpora; in fact, we could say that such corpora are almost non-existent. However, one recent, good effort in this domain is the KACST Text-To-Speech Database (KTD), an MSA simulated emotional speech read corpus produced by King Abdulaziz City for Science and Technology (KACST) (KACST, 2011). For this corpus, a professional actor simulated four emotions for all of the sentences. This corpus represented our reference in producing our new expanded and improved Arabic speech corpus, KSUKTD.

2.1. Design of Prompt Texts

We selected 16 sentences from the KTD corpus, as shown in Table 1. The KTD corpus contains many different types of sentences, but we selected only those sentences that simulated the four target emotions without any increases or decreases in word count as references for our speakers.

S1:	إِصَابَةٌ جَدِيدَةٌ بِشَلَلِ الْأَطْفَالْ، وَأَرْبَعُمِئَةَ وَخَمْسَةَ عَشَرَ بِالْجُذَامْ فِي الْيَمَنْ. ?isˀaabatun ʒadiidatun biʃalalili l?atˀfaal wa?arbaʕumi?ata waxamsata ʕaʃara bilʒuðaam fil jaman
S2:	أَلْعَدَدُ الْكُلِّيِّ لِمَرْضَى اَلْجُذَامْ، بَلَغَ مَعَ مَطْلَعِ الْعَامِ الْجَارِي، سَبْعَةَ آلَافٍ وَتِسْعِمِئَةٍ وَثَمَانِيَةً وَعِشْرِينَ حَالَةْ. ?alʕadadul kullijji limardˀal ʒuðaam balaʁa maʕa matˀlaʕil ʕaamil ʒaarii sabʕata ?aalaafin watisʕimi?atin waθamaanijatan waʕiʃriina ħaalah
S3:	وَفَاةُ الشَّيْخِ الْغَزَالِيّ، فِي الْمَدِينَةِ الْمُنَوَّرَةْ، فِي شَهْرِ مَارِسْ، عَامَ أَلْفٍ وَتِسْعِمِئَةٍ، وَسِتَّةٍ وَتِسْعِينْ. wafaatuʃ ʃajxil ʁazaalijj fil madiinatil munawwarah fii ʃahri maaris ʕaama ?alfin watisʕimi?atin wasittatin watisʕiin
S4:	وَفَاةُ اَلشَّيْخ جَادِ الْحَقْ، وَدَفْنُهُ فِي قَرْيَتِهِ بِطَرَّةْ، بِمَرْكَزِ طَلْخَا، بِالدَّقَهْلِيَّةْ. wafaatu ʃʃajx ʒaadi lħaq wadafnuhu fii qarjatihi bitˀurrah bimarkazi tˀalxaa biddaqahlijjah
S5:	أَلشَّيْخُ الشَّعْرَاوِيُّ، يُوَارَى الثَّرَى فِي دَقَادُوسْ. ?a ʃʃajxuʃ ʃaʕraawijj juwaaraθ θaraa fii daqaaduus
S6:	زَغْلُولِ النَّجَّارْ، يَتَكَلَّمُ عَنْ زِلْزَالِ تْسُونَامِي. zaʁluulin naʒʒaar jatakallamu ʕan zilzaalit suunaamii
S7:	أَحْمَدْ فُؤَادْ بَاشَا، عُضْوًا بِمَجْمَعِ الْخَالِدَيْن فِي اَلْقَاهِرَة. ?aħmad fu?aad baaʃaa ʕudˀwan bimaʒmaʕi lxaalidajn fi lqaahirah
S8:	بُورِسْ يِلْسِن، يَتَنَحَّى عَنِ السُّلْطَةِ لِأَسْبَابٍ صِحِّيَّةْ. buuris jilsin jatanaħħaa ʕanis sultˀati li?asbaabin sˀiħħijjah
S9:	جُورْجِ بُوش، يُقَدِّمُ وَسَاطَتَهُ لِحَلِّ الْأَزْمَةْ، بَيْنَ رُوسْيَا وَجُورْجِيَا. ʒuurʒi buuʃ juqaddimu wasaatˀatahu liħallil ?azmah bajna ruusjaa waʒuurʒijaa
S10:	مُحَمَّدْ رَجَبِ الْبَيُّومِي، رَئِيسُ تَحْرِيرِ مَجَلَّةِ الْأَزْهَرْ. muħammad raʒabil bajjuumii ra?iisu taħriiri maʒallatil ?azhar
S11:	أَلسَّادَاتْ، بَطَلُ اَلْحَرْبِ وَاَلسَّلَامْ. ?assaadaat batˀalul ħarbi wassalaam
S12:	كَامْبِ دِيفِيدْ، إِتِّفَاقِيَّةٌ مُلْزِمَةٌ بَيْنَ فِلَسَطِين وَإِسْرَائِيلْ. kaambi diifiid ?ittifaaqijjatun mulzimatun bajna filasatˀiin wa?israa?iil
S13:	أَلْعَاهِلُ الْمَغْرِبِيُّ، أَلْمَلِكُ اَلْحَسَنْ، فِي زِيَارَةٍ لِلْعَاهِلِ الْأَرْدُنِيِّ، أَلْمَلِكِ حُسَيْن بِن طَلَالْ. ?alʕaahilul maʁribijj ?almalikul ħasan fii zijaaratin lilʕaahilil ?ardunijj ?almaliki ħusajn bin tˀalaal
S14:	أَلْمَلِكْ فَهْدْ، يَتَوَعَّدُ تَنْظِيمَ الْقَاعِدَةِ فِي السُّعُودِيَّةْ. ?almalik fahd jatawaʕʕadu tanðˁiimal qaaʕidati fis suʕuudijjah
S15:	أَحْمَدْ هِيكَلْ، وَزِيرُ الثَّقَافَةِ الْأَسْبَقْ، يَحْصُلُ عَلَى جَائِزَةِ الدَّوْلَةِ اَلتَّقْدِيرِيَّةْ. ?aħmad hiikal waziiruθ θaqaafatil ?asbaq jaħsˀulu ʕalaa ʒaa?izatid dawlatit taqdiirijjah
S16:	بَشَّارِ الْأَسَدْ، وَلَحُّودْ، وَمُبَارَكْ فِي قِمَّةٍ ثُلَاثِيَّة. baʃʃaaril ?asad walaħħuud wamubaarak fii qimmatin θulaaθijjah

Table 1: Selected sentences

All 16 sentences were selected from a newspaper that can be accessed either visually or aurally through a variety of different media. In the case of the "questioning" emotion, we added the word **"هل /hal/**." In addition, for this emotion the shortest sentences contain four words, and the longest sentences contain 16 words. Figure 1 shows the phoneme frequency for all 16 sentences selected. Phoneme /a/ has the highest frequency, while phoneme /ðˁ/ has the lowest frequency. All Arabic phonemes are included, and their frequencies are relatively representative of Arabic, with /a/

and /l/ being the most frequent phonemes.

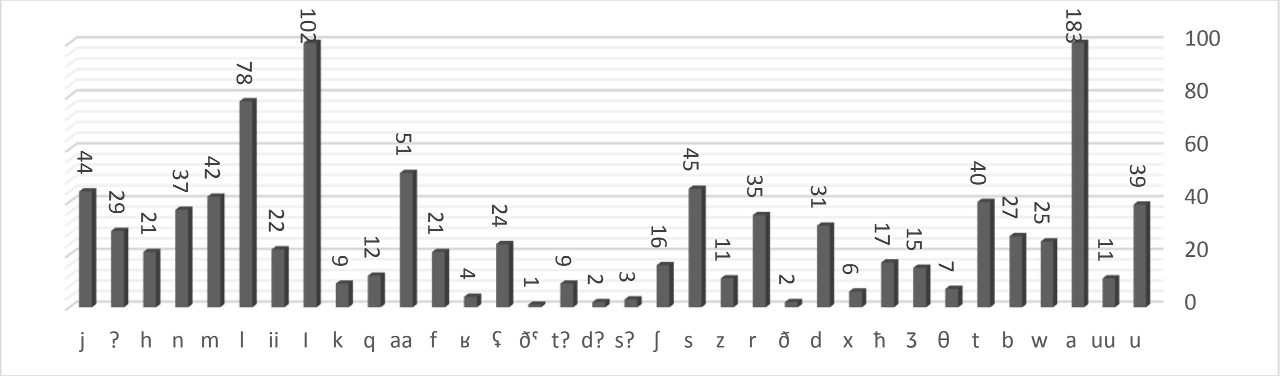


Figure 1: Phoneme frequency

Total number of words is 168 word; 115 are unique words and nine are repeated words, as shown in Table 2.

No. of Word	124
Repeated words	9
Non-repeated words	115
All words	152
“questioning” emotion sentences	16
Total	168

Table 2: Word statistics in KSUKTD corpus

2.2. Selection of Emotions

The first step in creating emotional speech corpora is the selection of emotions. One actor in the KTD corpus simulated the following four emotions: ***sadness, happy, surprised,*** and ***questioning.*** Our design for the new KSUKTD consisted of two phases. In the first phase, we selected the same emotions as those in the KTD corpus, with the addition of ***neutral speech***. Based on the results of an evaluation of this first phase, we moved to the second phase, as will be seen later in this paper.

2.3. Speaker Selection

Twenty male and female speakers recorded 16 sentences in the five different emotions. The speakers included 10 males aged between 20 and 37, and 10 females aged between 19 and 30. All speakers were either undergraduate or graduate students, except for one female who was still attending secondary school. Each speaker was asked to present an identification card that included his or her name, nationality, age, place of birth, location where part of his/her childhood was spent, where he/she currently lives, highest level of education achieved, educational levels of his/her parents, and marital status, etc.

2.4. Recording

The KSUKTD database was not recorded in a studio. Each person in charge of making the recordings was given the required devices to make the recordings, and then they traveled to each of the speakers’ homes so that the speakers could complete the recordings in their homes. The 20 speakers (10 male and 10 female) were from Saudi Arabia, Yemen, and Syria. High-quality microphones (SHURE 58 A) and three Dell laptops (model XPS 14Z) running Windows 7 were used to make the recordings. We used the Praat software program to control the mono recording processing. 16 KHz was the sampling frequency used. All 16 sentences were printed out for the speakers, and they were asked to read them many times before starting the recording.

2.5. Filename Format

The filename format of DxxExxPgxxSxxTxx was used, in which each file starts with “Dxx,” which indicates the corpus number. (This corpus is numbered 05, hence “D05.”) The next three digits, “Exx,” indicate the emotion code (E00, E01, etc.). This is followed by the code “Pgxx,” which indicates the speaker gender (0 male, 1 female) and number (001, 002, etc.). The code “Sxx” represents the sentence number (S01, S02,…S16), and, finally, “Txx” refers to the trial number (T01, T02, etc.). For example, D05E04P104S01T01 indicates that sentence 1 was recorded by female speaker number 4, who simulated the sentence using the “questioning” emotion. Table 3 shows the filename format details.

Dxx	**Exx**		**Pgxx**		**Sxx**		**Txx**	
Corpus	**Emotions**		**Persons**		**Sentences**		**Trials**	
05	E00	Normal	P0xx	Male	S01	Sentence#1	T01	Try # 1
	E01	Happy	P1xx	Female	S02	Sentence#2	T02	Try # 2
	E02	Sadness	P001	Ali	S03	Sentence#3	T03	Try # 3
	E03	Surprised	P101	Aisha	S04	Sentence#4	T04	Try # 4
	E04	Questioning			S05	Sentence#5	T05	Try # 5

Table 3: Filename format details

3. Perceptual Test

After the recordings were completed, we perform a perceptual test aiming at checking whether normal listeners could identify the recorded emotion types. For this test, we applied a set of rules for the listeners to follow. It is important to mention that no training session was provided before conducting the test in order to not influence the listeners. However, we allowed the listeners to ask the supervisor to stop at any time if they wanted to hear a recorded file again before deciding on the emotion type, but we did not allow them to go back and compare a recording with an earlier one spoken by the same speaker. Finally, the listeners could also take a break whenever they desired.

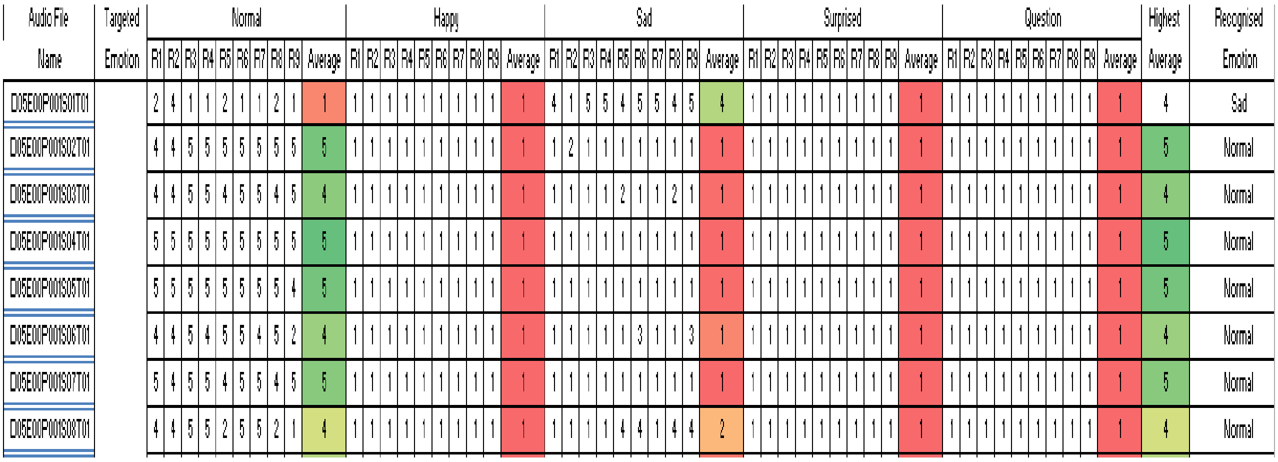
The test files were constructed as follows. First, all the filenames representing all of the recordings from all the speakers (1600 total files) were listed in two Excel spreadsheets and linked to their source audio files using hyperlinks. Then, the filenames were reordered randomly.. The nine listeners were six males and three females, all Arabic native speakers except for one male who was fluent in both written and spoken Arabic. All of the listeners were undergraduates in their 20s, except for the non-native listener, who was in his 40s. An exemple of the details for each listener's responses, including their ratings by percentage, are given in Table 4. The next step was to convert the collected data into mean opinion score (MOS), as shown in Table 5, by dividing all percentages by 20 and rounding off the result to obtain the final results shown in Table 6.

Listener No. 6							
File #	Normal (%)	Happy (%)	Sadness (%)	Surprised (%)	Questioning (%)	Male/Female	Notes
D05E00P105S08T01	80	0	20	0	0	F	
D05E01P006S07T01	30	70	0	0	0	M	noisy

Table 4: An example of Listener Table

MOS	Quality	Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

Table 5: Mean Opinion Score (MOS) (Ribeiro & Florêncio 2011)



Audio File	Targeted	Normal										Happy										Sad										Surprised										Question										Highest	Recognised
Name	Emotion	R1	R2	R3	R4	R5	R6	R7	R8	R9	Average	R1	R2	R3	R4	R5	R6	R7	R8	R9	Average	R1	R2	R3	R4	R5	R6	R7	R8	R9	Average	R1	R2	R3	R4	R5	R6	R7	R8	R9	Average	R1	R2	R3	R4	R5	R6	R7	R8	R9	Average	Average	Emotion
D05E00P001S01T01		2	4	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	4	1	5	5	4	5	5	4	5	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	Sad
D05E00P001S02T01		4	4	5	5	5	5	5	5	5	5	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	5	Normal
D05E00P001S03T01		4	4	5	5	4	5	5	4	5	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	Normal
D05E00P001S04T01		5	5	5	5	5	5	5	5	5	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	5	Normal
D05E00P001S05T01		5	5	5	5	5	5	5	5	4	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	5	Normal
D05E00P001S06T01		4	4	5	4	5	5	4	5	2	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	Normal
D05E00P001S07T01		5	4	5	5	4	5	5	4	5	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	5	Normal
D05E00P001S08T01		4	4	5	5	2	5	5	2	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	4	1	4	4	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	Normal

Table 6: Results in table format

4. Results and Analysis

Our goal is to perform human emotion classification of a recorded speech through listening sessions. The results of the listening test are shown in the following tables. The results are divided and listed in the following order: male and female results, speaker evaluation results, simulated emotion accuracy results, and sentence selection results.

4.1. Male Speakers

Table 7 shows the accuracy of each male speaker's results, ranked in order from highest to lowest accuracy with respect to the simulated emotions. From this table we can see that 60% of the speakers mastered the simulated emotions with accuracy greater than 80%, and that only one speaker had accuracy of less than 70%.

Male Speakers		
Speaker Number	**Best Simulated Emotions**	**Emotion Accuracy %**
P009	All	100
P001	Sad and Questioning	93.75
P005	Sad, Surprised and Questioning	93.75
P008	Sad and Surprised	88.75
P003	Normal, Surprised and Questioning	86.25
P004	Sad and Questioning	82.5
P006	Normal and Questioning	77.5
P002	Normal and Questioning	71.25
P010	Normal and Questioning	70
P007	Normal and Questioning	66.25

Table 7: Male speaker accuracy evaluations

Table 8 shows the percentage of correctly identified simulated emotions produced by the male speakers. The "questioning" emotion had the highest percentage of correct identification, at 98.75%, while the "happy" emotion had the lowest percentage of correct identification in the listening test evaluation, at 51.25%. The "neutral" and "surprised" emotions were both well simulated by the speakers and well identified by the listeners. We note, however, that the "sadness" emotion fell between the well (correctly) identified emotions (i.e., "questioning," "neutral," and "surprised") and the poorly (incorrectly) identified emotion of "happy."

Emotion	Correctly Recognized Emotions	Incorrectly Recognized Emotions	Percentage
Questioning	158	2	98.75
Neutral	151	9	94.375
Surprised	148	12	92.5
Sadness	126	34	78.75
Happy	82	78	51.25

Table 8: Male speaker emotion accuracy evaluations

In an attempt to determine whether there was an effect of the content of the sentence, Table 9 shows the sentences in order with respect to how well the target emotion was simulated by all the speakers. As shown in the table, sentence S16 has the highest score, while sentence S04 has the lowest score.

Sentence Number	Correctly Recognized Emotions	Incorrectly Recognized Emotions	Percentage
S16	46	4	92
S07	44	6	88
S13	44	6	88
S06	43	7	86
S02	42	8	84
S08	42	8	84
S12	42	8	84
S15	42	8	84
S09	41	9	82
S10	41	9	82
S11	41	9	82
S03	40	10	80
S05	40	10	80
S14	40	10	80
S01	39	11	78
S04	38	12	76

Table 9: Male speaker sentence emotion evaluations

4.2. Female Speakers

Table 10 shows the accuracy of each female speaker's results, ranked in order from highest to lowest accuracy with respect to the simulated emotions. From this table we can see that 40% of the speakers mastered the simulated emotions with accuracy greater than 90%, and that 50% of the speakers mastered the simulated emotions with an accuracy rate of 70%–77.5%. One female speaker had a low accuracy rate of 57.5%.

Female Speakers		
Speaker Number	**Best Simulated Emotion**	**Emotion Accuracy %**
P109	All	100
P101	Questioning	96.25
P107	Surprised	93.75
P110	Normal, Surprised and Questioning	92.5
P108	Sad and Questioning	77.5
P102	Normal, Sad and Questioning	75
P104	Questioning	73.75
P106	Surprised	73.75
P105	Normal	70
P103	Sad	57.5

Table 10: Female speaker accuracy evaluations

Table 11 shows the percentage of correctly identified simulated emotions produced by the female speakers. The "questioning" emotion had the highest percentage of correct identification, at 98.125%, while the "happy" emotion had the lowest percentage of correct identification in the listening test evaluation, at 56.25%. The "neutral" and "sadness" emotions were both well simulated by the speakers and well identified by the listeners. The "surprised" emotion evaluation fell between the well (correctly) identified emotions (i.e., "questioning," "neutral," and "sadness") and the poorly (incorrectly) identified emotion of "happy."

Emotion	Correctly Recognized Emotions	Incorrectly Recognized Emotions	Percentage
Questioning	157	3	98.125
Neutral	151	9	94.375
Sadness	133	27	83.125
Surprised	106	42	66.25
Happy	90	70	56.25

Table 11: Female speaker emotion accuracy evaluations

With regard to the selected sentences spoken by the female speakers, sentence S09 had the highest percentage of correct identification, while sentence S12 had the lowest percentage of correct identification as shown in Table 12

Sentence Number	Correctly Recognized Emotions	Incorrectly Recognized Emotions	Percentage
S09	43	7	86
S10	43	7	86
S11	43	7	86
S05	42	8	84
S13	42	8	84
S15	42	8	84
S03	41	9	82
S06	41	9	82
S08	41	9	82
S16	41	9	82
S02	39	11	78
S04	39	11	78
S14	39	11	78
S01	38	12	76
S07	38	12	76
S12	37	13	74

Table 12: Female speaker sentence emotion evaluations

In general, the male and female speakers' listening test results are approximately convergent, at 83.13% and 81.13%, respectively. The percentage of correctly identified emotion utterances across both genders was 82.13%, as shown in Table 13.

	Total Accuracy			
	Number of Files	Recognized	Unrecognized	Percentage
All	1600	1314	286	82.13%
Male	800	665	135	83.13%
Female	800	649	151	81.13%

Table 13: General listening test results

The "happy" emotion was the hardest for speakers to identify, with the successful identification rate not exceeding 53.75% and a difference of 27.2% from the next emotion, as shown in Table 14. In contrast, the "questioning" emotion was the most easily identifiable emotion by the listeners; this was followed by the "neutral," "surprised," and "sadness" emotions.

Emotion	Correctly Recognized Emotions	Incorrectly Recognized Emotions	Percentage
Questioning	315	5	98.4375
Neutral	302	18	94.375
Surprised	260	60	81.25
Sadness	259	61	80.9375
Happy	172	148	53.75

Table 14: Male and female speaker emotion accuracy evaluations

As for the sentence order, according to the number of sentences correctly recognized, as shown in Table 15, we can conclude that the reasons for this order may be due to the sentence length (e.g., S10, S02, and S03) and the fact that the sentence content may not have been commensurate with the emotion of "happy," (e.g., S10, S02, S03, and S04; all of these sentences include bad news about the death of someone or about the prevalence of disease). Thus, sentence content may have been a reason why the speakers failed to adequately simulate the emotion of "happy."

Male Female Sentences			
Sentence Number	Correctly Recognized Emotions	Incorrectly Recognized Emotions	Percentage
S16	87	13	87
S13	86	14	86
S15	84	16	84
S11	84	16	84
S10	84	16	84
S09	84	16	84
S06	84	16	84
S08	83	17	83
S07	82	18	82
S05	82	18	82
S03	81	19	81
S02	81	19	81
S14	79	21	79
S12	79	21	79
S04	77	23	77
S01	77	23	77

Table 15: Male and female speaker sentence emotion evaluations

5. Conclusion

An initial effort in designing, building, and presenting an MSA emotional speech database was attempted in this paper. We covered the design of prompt texts, the choice of target emotions, speaker selection, and database recording. In this effort, 10 male and 10 female speakers were asked to record 1600 files using a range of five emotions. The recorded corpus was perceptually tested by nine listeners (six male and three female). Based on the results of this listener perceptual test, the male speakers' simulated emotions were found to be more accurate than the female speakers' ones, and the emotion of "happy" was the least accurately identified emotion. In our future work, i.e., phase 2, which is already underway, we have added the emotion of "angry," selected only the best-performing speakers, and avoided sentences that have been deemed as having inappropriate content. In addition, we hope to expand the use of this corpus to include the analyzing and classifying of emotions of Arabic speakers by using the acoustic and prosodic features of speech signals as well.

6. Acknowledgment

This work was supported by the NPST program under King Saud University Project Number 10-INF1325-02.

7. References

Bhutekar, S., & Chandak, M. (2012). Designing and Recording Emotional Speech Databases. International Journal of Computer Applications (IJCA): Proceedings on National Conference on Innovative Paradigms in Engineering and Technology (NCIPET 2012), 4, 6–10.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., Berlin, T. U., … Berlin, H. U. (2005). A Database of German Emotional Speech. Interspeech., 5, 1517–1520.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), 572–587.

Engberg, I., & Hansen, A. (1997). Design, recording and verification of a danish emotional speech database. EUROSPEECH, ISCA, 1695–1698.

Hozjan, V., Kacic, Z., & Moreno, A. (2002). Interface Databases: Design and Collection of a Multilingual Emotional Speech Database. LREC, 2024–2028.

Jovi, S. T., Ka, Z., & Rajkovi, M. (2004). ISCA Archive Serbian emotional speech database : design , processing and evaluation. In 9th Conference Speech and Computer.

(KACST)., K. A. (2011). KTD Corpus. Riyadh: Unpublished Technical Report.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. International Journal of Speech Technology, 15(2), 99–117.

Kostoulas, T., Ganchev, T., Mporas, I., & Fakotakis, N. (2008). A Real-World Emotional Speech Corpus for Modern Greek. LREC, 2676–2680.

Navas, E., Castelruiz, A., Luengo, I., Sánchez, J., & Hernáez, I. (2004). Designing and Recording an Audiovisual Database of Emotional Speech in Basque. LREC, 1387–1390.

Ribeiro, F., & Florêncio, D. (2011). Crowdmos: An approach for crowdsourcing mean opinion score studies. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference On. IEEE, 2416–2419.

Saratxaga, I., & Navas, E. (2006). Designing and recording an emotional speech database for corpus based synthesis in Basque. In Proc. of Fifth International Conference on Language Resources and Evaluation (LREC), 2126–2129.

Staroniewicz, P., & Majewski, W. (2009). Polish Emotional Speech Database–Recording and Preliminary Validation. Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions. Springer Berlin Heidelberg, 42–49.

Large Arabic Web Corpora of High Quality: The Dimensions Time and Origin

Thomas Eckart, Faisal Alshargi, Uwe Quasthoff, Dirk Goldhahn
Natural Language Processing Group, University of Leipzig, Germany
Email: {teckart, alshargi, quasthoff, dgoldhahn}@informatik.uni-leipzig.de

Abstract

Large textual resources are the basis for a variety of applications in the field of corpus linguistic. For most languages spoken by large user groups a comprehensive set of these corpora are constantly generated and exploited. Unfortunately for modern Arabic there are still shortcomings that interfere with systematic text analysis. The use of the Arabic language in many countries with different cultural backgrounds and the political changes in many of these countries over the last years require a broad and steady text acquisition strategy to form a basis for extended analysis. This paper describes the Arabic part of the Leipzig Corpora Collection (LCC) which is a provider of freely available resources for more than 200 languages. The LCC focuses on providing modern text corpora and wordlists via web-based interfaces for the academic community. As an example for the exploitation of these resources it will be shown how wordlists reflect political and cultural concepts that can be automatically exploited for diachronic or spatial comparisons.

Keywords: Arabic corpus generation, text acquisition, comparative corpus analysis

1. Availability of Arabic Text Resources

For a language with such a large group of native speakers Arabic has (compared with similar languages) still a strong demand for large corpora and tools. Existing corpora like the Al-Hayat Corpus (Roeck, 2002), the Corpus of Contemporary Arabic (CCA) or the An-Nahar News Paper Text Corpus are valuable resources and widely used. Unfortunately many of the existing corpora or resources lack properties that are strongly desirable for their use in the scientific context. These shortcomings contain problems with their availability (in some cases only by using very specific interfaces), lack of currentness (a problem especially when dealing with ongoing political developments), high costs or strict licences that permit reuse and data aggregation. As some of these problems can't be eliminated in general (like in the context of copyright and personality rights) it would be desirable to have more resources that can be used with as less restrictions as possible and that can be useful for further progress in the exploitation of Arbic corpora and other text-based resources.

2. Arabic Resources at the LCC

The *Leipzig Corpora Collection* (LCC) collects digital text material for more than 20 years. Starting with a focus on European languages it became apparent that a lot of the developed strategies and tools could be reused for other languages as well. Over the last years the used tool chain for text acquisition and text processing was adopted to deal with non-Latin scripts and especially Arabic resources were created and constantly improved.

2.1. Text Acquisition Strategies

The *Leipzig Corpora Collection* (Goldhahn et al., 2012) combines different strategies for collecting textual data from the WWW. The main goal is to ensure that corpora of large extent and high diversity concerning topics or genres can be created for specific languages. Especially a language like Arabic that is spoken in many countries requires a variety of approaches to achieve this objective.

2.1.1. Generic Web Crawling

A framework for massively parallel Web crawling is applied that utilizes the standard Web crawler and archiver *Heritrix*[1] of the Internet Archive. Among other enhancements, it was enriched with means for the automatic generation of crawling jobs.

Heritrix is used in several ways. On the one hand whole Top Level Domains are crawled. In this case a small list of domains of a country of interest is used as an input. Heritrix is then configured to follow links within this top-level domain (TLD). This has been conducted for several countries where Arabic is an official language.

On the other hand News sources are downloaded using the Heritrix based Web crawler. Basis is a list of more than 32,000 news sources in about 120 languages provided by *ABYZ News Links*[2]. This service offers URLs and information regarding country and language. This way news texts for several Arabic countries were collected. This includes text data excluded in the TLD crawling because of non-country TLDs used such as ".com".

[1] http://webarchive.jira.com/wiki/display/Heritrix/Heritrix

[2] http://www.abyznewslinks.com

2.1.2. Distributed Web Crawling

FindLinks (Heyer and Quasthoff, 2004) is a distributed Web crawler using a client-server architecture. The Java-based client runs on standard PCs and processes a list of URLs, which it receives from the *FindLinks*-server.
FindLinks has been used with community support for several years and allowed us to crawl the WWW to a large extent.

2.1.3. Bootstrapping Corpora

In addition an approach similar to Baroni (2004) and Sharoff (2006) was applied. Frequent terms of Arabic are combined to form Google search queries and retrieve the resulting URLs as basis for the default crawling system.
A small set of frequent terms is needed for languages in question. Therefore existing corpora of the LCC or other sources such as the *Universal Declaration of Human Rights* (*UDHR*)[3] were utilized as a resource.
Based on these lists tuples of three to five high frequent words are generated. These tuples are then used to query Google and to collect the retrieved URLs, which are then downloaded.

2.1.4. Crawling of special Domains

Certain domains are beneficial sources for Web corpora since they contain a large amount of text in predefined languages.
One example is the free Internet encyclopedia Wikipedia, which is available in more than 200 languages and of course also contains a version in Arabic.
Wikipedia dumps for these languages, among them Arabic, were downloaded. *Wikipedia Preprocessor*[4] was used for further processing and text extraction.

2.2. Corpus Creation Toolchain

Necessary steps for the creation of dictionaries are text extraction (mostly based on HTML as input material), language identification (Pollmächer, 2011), sentence segmentation, cleaning, sentence scrambling, conversion into a text database and statistical evaluation.
An automatic and mainly language independent tool chain has been implemented. It is easily configurable and only few language-dependent adjustments, concerning e.g. abbreviations or sentence boundaries, have to be made.
In a final step statistics-based quality assurance is applied to achieve a satisfying quality of the resulting dictionaries (Quasthoff, 2006b) (Eckart, 2012). Using features such as character statistics, typical length distributions, typical character or n-gram distributions, or tests for conformity to well-known empirical language laws problems during corpora creation can be detected and corrected.

The processing of Arabic text required several changes to the existing toolchain. Most of the developed tools could be reused but specific configurations had to be changed. This includes changes to components like sentence segmentation or quality assurance procedures. Besides some minor problems the general system again proved to be stable enough as for other languages or scripts before.

2.3. Sentence Scrambling

For all corpora the sentences had to be "scrambled" to destroy the original structure of the documents due to copyright restrictions. This inhibits the reconstruction of the original documents. With respect to German copyright legislation this approach is considered safe.

2.4. Available Resources

Corpora of this collection are typically grouped regarding the dimensions language, country of origin, text type (newspaper text, governmental text, generic Web material, religious texts etc.) and time of acquisition. The following table gives an introduction of currently available resources. As the crawling is an ongoing process new corpora are added at least every year.
Currently there are country specific corpora for Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Libanon, Mauritania, Morocco, Oman, Palestine, Qatar, Sudan, Syria, Tunisia, United Arab Emirates and Yemen which mostly consist of newspaper articles. As an example table 1 shows the most frequent sources of input material for a Morocco Web corpus.

Domain	*Number of documents*
www.riyada.ma	40,488
bayanealyaoume.press.ma	40,005
www.aktab.ma	39,309
www.goud.ma/	37,270
www.almassae.ma	37,203
www.almassae.press.ma	33,371
www.attarikh-alarabi.ma	33,255

Table 1: Number of documents for the most frequent sources used for a Moroccan Web corpus from 2013

2.5. Available Interfaces

The corpora are available via different Web-based interfaces. There is a freely available web portal where a variety of information can be accessed based on a word level (like sample sentences, word co-coocurrences, co-co-occurrence graphs etc.)[5]. Furthermore many corpora can be downloaded for free in different formats. These include plain text versions of the textual material and also MySQL databases[6]. For the later the platform-independent

[3] http://www.ohchr.org

[4] http://sourceforge.net/projects/wikiprep/

[5] Arabic portal: http://wortschatz.uni-leipzig.de/ws_ara/

[6] http://corpora.informatik.uni-leipzig.de/download.html

browsing tool is provided which allows examining the corpus locally.

3.

3.1. Linguistic Variants and Spell Checking

There are large sets of linguistic variants for many Arabic terms in the corpora. This is due to different reasons: there are many Arabic dialects spoken in different countries (like the term خمسا (*five)* which is used in Saudi Arabia, or خمسـتلاف(*five thousand)* which is used in Egypt). Besides these regional specifics there are of course also a lot of spelling errors like خمسنئة (*Five hundred*).

Table 2 gives a short impression of different variants of the same word including their word rank in a Arabic mixed corpus with more than 4 million sentences.

3.2. Diachronic Comparisons

The availability of diachronic corpora can be used to detect political, economic and even cultural changes. These changes directly reflect in journalistic texts and user generated content.

Table 3 shows an example of such a diachronic comparison. The word rank of several terms being used in political contexts are calculated for six newspaper corpora based on input material from several Arabic speaking countries for the years 2007 to 2012. As expected words being part of current controversial topics are subject of strong changes in their relative frequency which is reflected in their word class. As an example *Obama* does hardly occur before 2008, but has a dramatically increase in frequency over the next years, with its peak in 2009 with the election of Barack Obama as US president in January.

3.3. Comparisons between Countries and Regions

By using texts from different top level domains it is furthermore possible to compare the contextual use of words in different countries. Based on sentence co-occurrences the generated co-occurrences graphs directly reflect typical usage of a word in a country and hence political situation and opinions. By comparing these graphs it is possible to extract similarities and differences in the public perception of different kind of topics.

Figure 1 shows the typical contexts of the word الانتخابات (Election) for text corpora from Bahrain and Egypt from 2013. Apparently some of the co-occurring terms are the same for both corpora (like *parliament*, *politics, voting* and similar election-related terms). However there are also differences: in Bahrain we also see the term *women.* This is because of the novelty of women allowed to vote in elections in Bahrain. On the other side both graphs contain different words for *vote:* الاقتراع in Bahrain and its Egypt correspondent التصويت.

4. Outlook

This corpora collection will continue in aggregating Web-based text material to extend the amount and quality of available resources. The result of these efforts will be furthermore provided to all interested users. Until mid of 2014 a new Web portal will be deployed that provides extended functionality and a more user-friendly interface. The underlying RESTful web services are also openly available and can be used for external applications as well. As a next step in exploiting word lists as a valuable resource in information extraction and language comparison it is planned to publish a book in the series of frequency dictionaries focusing on word frequency information in the Arabic language.

5. References

Al-Sulaiti L., Atwell E. (2004), Designing and developing a corpus of Contemporary Arabic. In *Proceedings of the sixth TALC conference*. Granada, Spain.

Baroni, M., Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*.

De Roeck, A. (2002). ELRA's Al-Hayat Dataset: Text Resources in Arabic, Language Engineering. In *ELRA Newsletter 2002*, Vol.7 No.1.

Eckart, T.; Quasthoff, U.; Goldhahn, D. (2012). Language Statistics-Based Quality Assurance for Large Corpora. In *Proceedings of Asia Pacific Corpus Linguistics Conference 2012*, Auckland, New Zealand.

Goldhahn, D., Eckart, T., Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of LREC 2012* (pp. 759-765)

Heyer, G., Quasthoff, U. (2004). Calculating Communities by Link Analysis of URLs. In *Proceedings of IICS-04*, Guadalajara, Mexico and Springer LNCS 3473.

Pollmächer, J. (2011). Separierung mit FindLinks gecrawlter Texte nach Sprachen. Bachelor Thesis, University of Leipzig.

Quasthoff, U., Biemann, C. (2006). Measuring Monolinguality. In *Proceedings of LREC 2006 Workshop on Quality assurance and quality measurement for language and speech resources.*

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini, editors, WaCky! Working papers on the Web as Corpus. Gedit, Bologna.

Words	*Correct form*	*Translation in English*	*Rank in Wordlist*	*Comment*
خمة	خمسة	Five	751,951	Spelling error
خمسا	خمسة	Five	85,625	Used in Saudi dialect
خمس	خمسة	Five	1,122	All Arabic MSA and Dialects
خمسائة	خمسمائة	Five hundred	359,873	Spelling error
خمستعشر	خمسة عشر	Fifteen	1,438,010	Used in Yemen dialect
خمستلاف	خمسة آلاف	Five thousand	1,438,011	Used in Egypt dialect
خمسنئة	خمسمائة	Five hundred	1,438,019	Spelling error
خمسيه	خمسمائة	Five hundred	751,973	Used in Jordan dialect

Table 2: Examples for language variants and spelling errors in MSA

English	*Term*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*
Democracy	الديمقراطية	631	721	453	655	347	500
Israel	إسرائيل	168	118	88	99	114	195
Obama	أوباما	10485	173	93	195	187	630
Elections	الانتخابات	170	141	97	153	138	158
Rights	الحقوق	683	2063	1180	1590	2892	1507
Iran	إيران	141	190	104	147	215	291
Freedom	الحريه	1635	1372	1175	656	699	636
Gaddafi	القذافي	1959	1894	2134	3804	79	589
Brotherhood	الاخوان	6556	5763	23147	15725	5122	2895

Table 3: Word rank of different terms in Arabic newspaper corpora from 2007 to 2012

Bahrain newspaper corpus 2013	*Egypt newspaper corpus 2013*
الانتخابات - alantxabat - Elections	الانتخابات - alantxabat - Elections

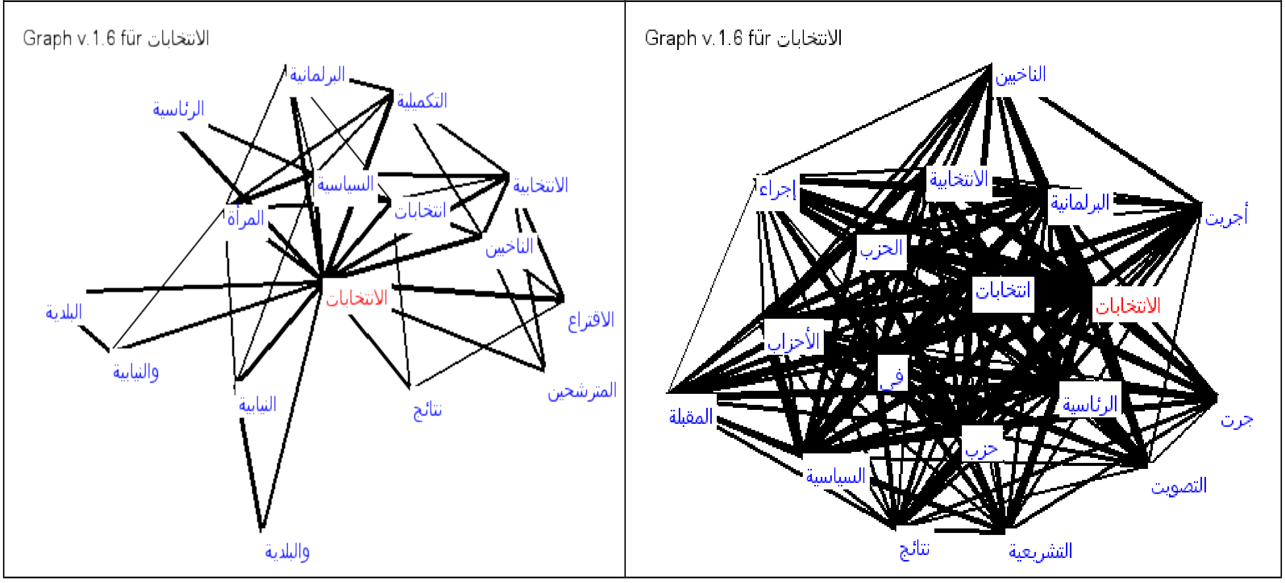


Figure 1: Word co-occurrences graphs of two newspaper corpora based on material from Bahrain and Egypt in 2013

An Algerian Arabic-French Code-Switched Corpus

Ryan Cotterell,[1] Adithya Renduchintala,[1] Naomi Saphra,[1] Chris Callison-Burch[2]

[1] Center for Language and Speech Processing, Johns Hopkins University
[2] Computer and Information Science Department, University of Pennsylvania

Abstract

Arabic is not just one language, but rather a collection of dialects in addition to Modern Standard Arabic (MSA). While MSA is used in formal situations, dialects are the language of every day life. Until recently, there was very little dialectal Arabic in written form. With the advent of social-media, however, the landscape has changed. We provide the first romanized code-switched Algerian Arabic-French corpus annotated for word-level language id. We review the history and sociological factors that make the linguistic situation in Algerian unique and highlight the value of this corpus to the natural language processing and linguistics communities. To build this corpus, we crawled an Algerian newspaper and extracted the comments from the news story. We discuss the informal nature of the language in the corpus and the challenges it will present. Additionally, we provide a preliminary analysis of the corpus. We then discuss some potential uses of our corpus of interest to the computational linguistics community.

Keywords: code-switching, Algerian Arabic, romanized Arabic, French

1. Introduction

Language identification systems have long operated under the assumption that text is written in a single language. As social media becomes a more prominent mode of communication, such systems are confronting text that increasingly challenges the monolingual assumption. More than half the world's population is bilingual and information communication is often code-switched, reflecting the need for a deeper understanding of code-switching in relation to NLP tasks. Recent work has proposed both supervised and unsupervised methods for *word*-level language id. Current methods, however, rely on the assumption that external resources exist, such as large non-code-switched corpora and dictionaries. These resources are not available for some languages and dialects, including Algerian vernacular Arabic, a dialect often code-switched with French. We are releasing a corpus of romanized Algerian Arabic and French scraped from the comments sections of Echorouk, an Algerian daily newspaper with the second-largest reader base of any Arabic paper. This is the only significant corpus of romanized Arabic known to the authors and additionally it is the largest corpus of code-switched data to our knowledge. Such a resource is necessary because romanized Arabic is becoming increasingly popular on the internet. The corpus will also be of use for the linguistic study of code-switching. Much of previous code-switching research has focused on data collected from field work, and a found dataset like ours could provide an interesting perspective on the use of code-switching in conversation.

2. Code-switching

Code-switching is a linguistic phenomenon wherein speakers switch between two or more languages in conversation, often within a single utterance (Bullock and Toribio, 2009). It can be viewed through a sociolinguistic lens where situation and topic influence the choice of language (Kachru, 1977). To define code-switching as a phenomenon, it is important to make the distinction between *code-switching* and *borrowing*. *Borrowing* is the act of using a foreign word without recourse to syntactic or morphological properties of that language and often occurs with phonological assimilation. *Code-switching*, on the other hand, involves switching between languages in which the speakers are fluent, and can in effect be viewed as changing the grammar in use. Some linguists have even proposed a scale of code-switching, positing the existence of a continuum between *borrowing* and *code switching* (Auer, 1999). Code-switching points (the times at which speakers change language) and the context around which switching occurs are also of interest to linguists. These points often lie within a sentence and their position is influenced by the syntax of the respective languages. Poplack (Poplack, 1988) posited that code-switching points cannot occur within a constituent. Recent work, however, has found that many speakers relax this constraint. The Matrix Language-Frame (MLF) model is one theory has gained traction (Myers-Scotton, 1993) to explain code-switching patterns. MLF proposes that there is a Matrix Language (ML) and an Embedded Language (EL). The ML is the more dominant language and is often the language which the speaker identifies as their native tongue. The EL is then inserted into the ML at certain grammatical frames. Within this framework, further work has gone into the exact syntactic and morphological contexts that allow for code-switching points (Myers-Scotton and Bolonyai, 2001).

3. Code-Switching in North Africa

Code-switching in North African Arabic is an established phenomenon that has been studied by the linguistics community (Bentahila and Davies, 1983). It dates back to the initial French colonization of North Africa. North Africa is also home to many cultures, a fact which potentially affects language use and code switching in particular. Until recently, mixed language communication has been observed mainly as a spoken phenomenon. With the widespread use of computer-mediated communication, code-switching is becoming common in North African Arabic (Salia, 2011).

Comments on news-feeds and social media outlets like Twitter and Facebook often contain code-switching. North African Arabic is not the only language to appear in code-switching writing. A body of recent sociolinguistics work has considered the phenomenon in various settings. Swiss-German and German code-switching in chat rooms was analyzed in Siebenhaar (2006) and Callahan (2004) considered Spanish and English code-switching in a written corpus.

4. Related Work

From a computational perspective, code-switching has received relatively little attention. Joshi (1982) provides a tool for parsing mixed sentences. More recently, Rosner and Farrugia (2007) focused on processing code-switched SMS. Solorio and Liu (2008) trained classifiers to predict code-switching points in Spanish and English. Nguyen and Dogruoz (2013) also focused on word level language identification in Dutch-Turkish news commentary. To our knowledge, Elfardy and Diab (2012) is the only computational work on Arabic code-switching done to date. That work does not include romanized Arabic. Many languages written in a non-Roman script are *romanized* on the internet. This practice presents a problem for standard NLP tools that are trained on the language with its standard orthography (Irvine et al., 2012). We believe that this type of romanized data will become more pervasive as more users employ computer-mediated communication globally. More information will be generated in such settings, and it is critical for future NLP systems to be able to process the data produced. The corpus we are presenting is a step in this direction.

5. Data Collection

We used a corpus crawled from an Algerian newspaper website. We scraped 598,047 pages in September 2012. These fora are rich in both dialectal Arabic and French content. The corpus contains discussion on a wide-ranging set of issues including domestic politics, international relations, religion, and sporting events. We extracted 6,949 comments, containing 150,000 words in total. We separated the comments section from the main article on each page and stripped HTML tags and other non-user generated content. The metadata was stripped in an attempt to preserve anonymity. The Arabic portion of the corpus was annotated for sentence level dialect on Mechanical Turk (Cotterell and Callison-Burch, 2014).

We separated all the comments, in which more than half the non-white space characters were in the Roman alphabet, determining these to be romanized. We did no further processing, e.g. tokenization. The final data set contains 339,504 comments with an average length of 19 tokens, as determined by separating on white space and punctuation. 1,000 of the comments are annotated using the guidelines described below. Our corpus has 493,038 types and 6,718,502 tokens, and is formatted in JSON.

6. Romanization of Arabic

This corpus is unique in that it is the first large corpus to the authors' knowledge that is composed of of an Arabic dialect written in romanized form. Romanized Arabic is particularly difficult because there is no standard form of romanization used across the Arab world. In order to use standard NLP tools on such corpora, it is often necessary to *deromanize* the corpus. In the case of Urdu, this task has been successfully completed using standard Machine Translation software (Irvine et al., 2012).

Arabic	Arabizi	Arabic	Arabizi
ا	a	ب	b, p
ت	t	ث	th, s
ج	j, g	ح	7, h
خ	7', 5	د	d
ذ	th z	ر	r
ز	z	س	s c
ش	sh, ch	ص	9
ض	9', d	ط	t
ظ	th	ع	3
غ	gh, 3'	ف	f
ق	8, 2, k, q	ك	k
ل	l	م	m
ن	n	ه	h
و	w, o, ou	ي	y, i, e

Figure 1: Correspondence Between Arabic Letters and Romanized Arabic (Yaghan, 2008)

Arabic written in the Latin alphabet, often dubbed *arabizi*, is extremely common on the internet and SMS. The exact mapping from the Arabic script onto the Latin alphabet varies significantly between regions. The specific case of romanization by young speakers of Gulf Arabic in the United Arab Emirates is discussed thoroughly in Palfreyman and Khalil (2003).

Figure 1 expresses the most common mappings across the Arab world. Algerians, and North Africans in general, tend to use romanizations that reflect French orthography: for instance و ↦ ou, ش ↦ ch, ج ↦ dj and ا ↦ è or é. To illustrate this difference consider the frequency of the common transcriptions of إن شاء الله (God willing); we see RL ↦ ch about an order of magnitude more often than ش ↦ sh.

This transcription variation makes it unlikely that a single, general-purpose Arabic deromanization tool will be enough, and such romanized corpora will need to be developed for other dialects as well in order to analyze the users romanization preferences on dialectal basis.

7. Text Analysis

Because our Algerian corpus is from the length-constrained informal domain of online forum comments, it would be difficult to process meaningfully without normalizing beforehand. It exhibits extreme variation in spelling and grammar. Many forms of the same word may appear throughout our corpus. For example, we identified 69 variants of the common word إن شاء الله alone. 70% of all token types in our corpus appeared only once, so the OOV rate in this forum corpus can confound language processing systems without text normalization. Several typical sources of variation for Arabic identified by (Darwish et al., 2012) were found in our corpus.

- The use of elongations, especially in the form of vowel repetition.

  > we ki ta3arfou wach rah testfadou ? **hhhhh** cha3ab **kar3adjiiiiiiiii**

- Spelling mistakes, such as dropped or transposed characters.
- Abbreviations.

  > rah thablona bel **BAC** had al3am !!!!!

- Emotional tokens and ejaculative abbreviations, such as the abbreviation “lol” borrowed from English web speech, or emoticons.

In addition to these irregularities, our corpus contains variations particular to romanized Arabic text because there is no standardized way to transcribe Arabic orthography in this informal domain, Arabic words can be represented by multiple spellings.

8. Example Posts

We present below a few example sentences that we have collected use the data collection methodology described above.

- bezaf m3a saifi oalah mnkalifoha mairbahch
  Had enough with Saifi.
- la howla wa la kowata il bi lah el3alier l3adim wa la yassa3oni an akoul anaho kllo chaye momkin ma3a ljazairyine
  For Gods Sake! I can just say that anything is possible with Algerians .
- we ki ta3arfou wach rah testfadou ? hhhhh cha3ab kar3adjiiiiiiiii
  Don’t try to know everything because it does not matter to you.
- 7ade said nchalah li lmontakhabina el3assekari nchalah yjibo natija mli7a bitawfiiiiiiiiiiiiiiik . . . onchoriya chorouk stp
  Good luck to our military team, I hope they get a good score. Good Luck! Say it chorouk!
- ya khawti tt simplement c bajiou l3arab o makanech fi tzayer kamla joueur kima lhadji et on vai ras le 27 03 2011 chkoun houma rjal liyestahlouha
  My brothers he’s simply the Baggio of Arabs, and there is none like Lhadji in Algeria, and on 27th of march 2011 we will see who wins.
- mais les filles ta3na ysedkou n’import quoi ana hada face book jamais cheftou khlah kalbi
  Our girls believe anything, I have never seen this Facebook before.

9. Annotation Guidelines

Annotating for word level language identification in code-switched text is a difficult task because whether a word is code-switched is often more of a continuum than a binary decision. Place names form a simple example: باريس (Paris) is an MSA word in that it is found in most Arabic dictionaries, but it is clearly of French origin. On the other hand, فيديو (video) is an example of a recent borrowing from European languages that should be considered an Arabic word. To simplify the decision, we made use of guidelines for dialectal Arabic annotation provided in (Elfardy and Diab, 2012). Their guidelines were created for annotating world level language id in a corpus composed of mixed dialectal Arabic and MSA both written in the Arabic script. As we annotating two linguistically dissimilar languages, the same level of ambiguity does not arise.

Further research in area of code-switching should focus on richer annotation schemata that are both linguistically motivated, i.e. taking into account the continuum of code-switching, and serve practical NLP needs. Another interesting area to focus on could be to annotate broad categories describing the type of code-switching. Kecskes (2006) describes three prominent patterns in code-switching (Insertion, Alternation and Congruent Lexicalization) based on Gibraltar data. Insertion involves adding lexical items from one language into the structure of the other. Alternation is similar to insertion except that larger chunks are inserted, rather than single tokens. Congruent Lexicalization is adding lexical items from different lexical inventories into a common grammar structure. The dataset could annotate each point of code-switching with these patterns of code-switching as well. This data set, however, is focused on only the points of code-switching.

The annotators were presented with posts and asked to label each word, split on white space and punctuation. They were given the choice of Arabic (A), French (F) and Other (O). We excluded punctuation from the annotation. Figure 4 shows the distribution of these tags in our dataset. The annotation was conducted with an interactive Python script.

10. Potential Uses

The corpus provided is the first of its kind in that it is the first large corpus of romanized Arabic that is code-switched with another language. This has numerous potential uses in both the NLP community and the linguistics community. In the NLP community, the processing of informal text is becoming and increasingly popular task among researchers (Yang and Eisenstein, 2013). This corpus adds another complication to informal text processing with the addition of code-switching. In the linguistics community, a corpus based analysis of a code-switched corpus offers the possibility to test various hypotheses on large number of documents. The MLF hypothesis has already been studied in bilingual speech corpora from Miami, Patagonia, and Wales (Carter, 2010), but it will be necessary to study such theories in the context of many distinct languages and cultures to gain deeper insight into the code-switching phenomenon.

11. Conclusion and Future Work

The primary contribution of this paper is the release of a Algerian Arabic-French code-switched corpus. We have made use of a previously proposed annotation scheme for word level language identification and highlighted the unusual qualities of this corpus that make it a significant contribution to field. Future work in this line should largely focus on experiments using the corpus both in NLP and linguistics. It would also be of interest to construct and annotate similar corpora for other informal code-switched Arabic dialects.

Acknowledgements

This material is based on research sponsored by a DARPA Computer Science Study Panel phase 3 award entitled "Crowdsourcing Translation" (contract D12PC00368). The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements by DARPA or the U.S. Government. This research was supported by the Johns Hopkins University Human Language Technology Center of Excellence. Adithya Renduchintala acknowledges support of NSF (IIS-1349902).

12. References

Peter Auer. 1999. From codeswitching via language mixing to fused lects toward a dynamic typology of bilingual speech. *International Journal of Bilingualism*, 3(4):309–332.

Abdelali Bentahila and Eirlys E Davies. 1983. The syntax of Arabic-French code-switching. *Lingua*, 59(4):301–330.

Barbara E Bullock and Almeida Jacqueline Toribio. 2009. *The Cambridge handbook of linguistic code-switching*. Cambridge University Press Cambridge.

Laura Callahan. 2004. *Spanish/English codeswitching in a written corpus*, volume 27. John Benjamins Publishing.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written arabic. In *LREC*.

Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for Arabic microblog retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2427–2430, New York, NY, USA. ACM.

Heba Elfardy and Mona T Diab. 2012. Token level identification of linguistic code switching. In *COLING (Posters)*, pages 287–296.

Ann Irvine, Jonathan Weese, and Chris Callison-Burch. 2012. Processing informal, Romanized Pakistani text messages. In *Proceedings of the Second Workshop on Language in Social Media*, pages 75–78. Association for Computational Linguistics.

Aravind K Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th conference on Computational linguistics-Volume 1*, pages 145–150. Academia Praha.

Braj B Kachru. 1977. Linguistic schizophrenia and language ensus: A note on the Indian situation. *Linguistics*, 15(186):17–32.

I Kecskes. 2006. A dual language model to explain code-switching: A cognitive-pragmatic approach. *Intercultural Pragmatics*, 3:257–283.

Carol Myers-Scotton and Agnes Bolonyai. 2001. Calculating speakers: Codeswitching in a rational choice model. *Language in Society*, 30(1):1–28.

Carol Myers-Scotton. 1993. Common and uncommon ground: Social and structural factors in codeswitching. *Language in Society*, 22:475–475.

Dong-Phuong Nguyen and AS Dogruoz. 2013. Word level language identification in online multilingual communication. Association for Computational Linguistics.

David Palfreyman and Muhamed al Khalil. 2003. a funky language for teenzz to use: representing gulf Arabic in instant messaging. *Journal of Computer-Mediated Communication*, 9(1):0–0.

Shana Poplack. 1988. Contrasting patterns of code-switching in two communities. *Codeswitching: Anthropological and Sociolinguistic Perspectives. New York: Mouton de Gruyter*, pages 215–244.

Mike Rosner and Paulseph-John Farrugia. 2007. A tagging algorithm for mixed language identification in a noisy domain. In *INTERSPEECH*, pages 190–193.

R. Salia. 2011. *Between Arabic and French Lies the Dialect: Moroccan Code-Weaving on Facebook*. Undergraduate thesis, Columbia University.

Beat Siebenhaar. 2006. Code choice and code-switching in Swiss-German Internet Relay Chat rooms. *Journal of Sociolinguistics*, 10(4):481–506.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 973–981. Association for Computational Linguistics.

Mohammad Ali Yaghan. 2008. Arabizi: A contemporary style of Arabic slang. *Design Issues*, 24(2):39–52.

Yi Yang and Jacob Eisenstein. 2013. A log-linear model for unsupervised text normalization. In *Proc. of EMNLP*.

An Open Platform Based on HPSG Formalism for the Standard Arabic Language

Mourad Loukam (1), Amar Balla (2), Mohamed Tayeb Laskri (3)
(1) Natural and Arabic Language Processing Team , Mathematics and its applications Laboratory, Faculty of Sciences,
Hassiba Benbouali University of Chlef, Algeria
(2) High Computer School, Algiers, Algeria
(3) Department of Computer Science, Faculty of Sciences , Badji Mokhtar University of Annaba, Algeria
E-mail: mourad.loukam@univ-chlef.dz, a_balla@esi.dz, laskri@univ-annaba.org

Abstract

HPSG formalism knows since many years a great development in NLP. We are working on the HPSG formalism on the double aspect of modelling / implementation with the aim of its application to the standard Arabic language. In this paper, we present an open platform, based on HPSG formalism, for the standard Arabic language. The originality of the platform is that it is an integrated tool which offers the complete chain of parsing texts in Arabic language in order to produce their HPSG analysis. In the medium-term, our objective is to use the platform for developing applications for Arabic NLP.

Keywords: Arabic Language, HPSG, NLP, Platform.
.

1. Introduction

The Head driven Phrase Structure Grammar (HPSG) formalism , was introduced by Carl Pollard and Ivan Sag in 1993 (Pollard & Sag, 1993). It is undoubtedly one of the most prominent theories currently in use in natural language processing (NLP) researches.

Several features make HPSG attractive for researchers, in particular:

- HPSG inherits from earlier formalisms (GPSG , TAG , LFG ...) retaining their advantages and avoiding their disadvantages.
- HPSG opts for the richness of linguistic units' representation. Morphological, syntactic and even semantic information are represented together in a single and homogenous structure referred to as the feature structure.
- For HPSG, the analysis process is based on the unification theory, which is well known in programming languages such as Lisp and Prolog.
- HPSG uses a reduced set of rules that can be applied, in theory, to all languages .
- HPSG seems well suited for computer processing, which it replicates directly many concepts such as typing and inheritance. .

The analysis of a sentence in HPSG consists of finding "the head" (the rector or dominant element), which then guides the analysis to the detection of the other elements of the sentence. Often, it is a real challenge to identify the head of a sentence, especially for sentences containing complex structure such as passive, interrogative, relative, and coordination structures.

We are working on the HPSG formalism on the double aspect of "modelling" and "implementation" for its application to Standard Arabic language (Loukam & al., 2013).

In this paper, we present the Pharas project (a Platform based on HPSG formalism for analysing Arabic Standard texts). Pharas is an integrated tool that offers the whole chain for analyzing Arabic texts, in order to produce their representation in HPSG format. Currently, the platform is operational including several modules: a subsystem for morpho- lexical analysis, a parser using unification and many resources (verbs , nouns , proper names, particles and dictionaries).

2. Related Works

Works on HPSG can be classified into two categories: those in the area of modelling and those focusing on the implementation.

The modeling aims to provide an analysis of a given linguistic phenomenon (passive sentence, interrogative, relative, coordination, …). Many works of this class can be found in the Proceedings of the Annual Symposium HPSG (http://hpsg.stanford.edu/) and as an example we can cite, Hann, 2011, 2012).

The implementation aims to develop tools and applications to produce automatic analysis using HPSG concepts. Some of these tools include:

- LKB (Linguistic Knowledge Building) is a system of grammatical development created by Ann Copestake and her team at the University of Cambridge (Copestake, 2002). This tool was not designed specifically for HPSG grammars, but it is a development platform for implementation of grammars based on unification.
- Trale : is a platform for implementing HPSG grammars , derived from the Milcah project developed at the University of Bremen (Germany). It has been applied to the German language for teaching theoretical linguistics.
- Matrix :is an experimental platform , supported by nearly a dozen research laboratories. Its goal is to offer an environment for rapid development of new grammars. But, the most important idea of this project is to design a universal grammatical core giving signature base (general types , simple lexical types, combination rules) and a set of parameterized modules (questions, negation, coordination, …) which can be used, in theory, to any

natural language (Emely and Lascarides, 2013).
Enju : is an HPSG parser for English, developed by Tsujii laboratory of Tokyo University. It has a wide coverage of the English grammar and has been tested on probabilistic models, especially in the biomedical field (Ninomiya & al., 2006), (Miyao & Tsujji, 2005).
Babel : an HPSG parser for German, developed at Berlin University (Muller, 2001).
With regard to the Arabic language processing using HPSG, very little work of this sort has been done. We can cite Maspar system developed at the University of Sfax/Tunisia (Bahou & al., 2006).

3. The Pharas Project

The Pharas project has been launched after it was noticed that existing systems using HPSG formalism are not well suited for the standard Arabic language processing.
The main idea was to design and to develop an integrated environment using HPSG formalism and containing the whole chain of parsing modules, in order to produce the analysis result of any given standard Arabic text. In addition to this experimental aspect, the aim is, in the medium-term, to come up with a platform which can be used for developing applications for the Arabic language.

From the beginning, we have adopted two orientations related to the development strategy :
– Extensibility: The platform will be developed using a modular approach to allow to integration of other NLP modules and also the use of API (Application programme Interface) when used for other applications using the Arabic language. The platform resources will be made available for external use. .
– open- source : preference has been given to open-source development tools and languages, such as Java, XML.

3.1 The General Analysis Process

Arabic texts introduced to Pharas platform for analysis will follow several steps as shown in Figure 1.

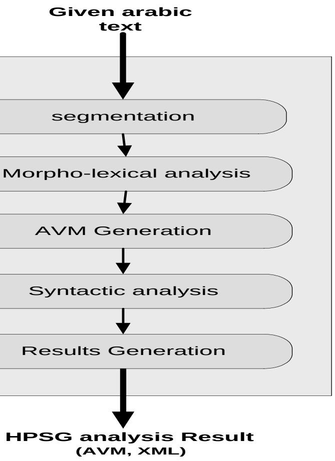


Figure 1: General analysis process in Pharas

We can summarize this process as follows:
1 . the given text is segmented into "words".
2 . each word obtained from the previous step is submitted to the morpho- lexical module wich examines each possible form the word can have (simple noun, verb, particle, derivated or flected form, …).
3 . if the word is recognized, a set of attribute/value matrix (AVM) for the word will be generated.
4 . the syntactic parser module operates on the results obtained from the morpho-lexical phase in order to grammaticality check the sentence. The parser uses the system rules and the unification process .
5 . The final result is given in attribute/value matrix (AVM) or XML format.

3.2 Architecture

The architecture of Pharas platform is based on the interconnection of multiple subsystems requiring various resources and tools (Figure 2).

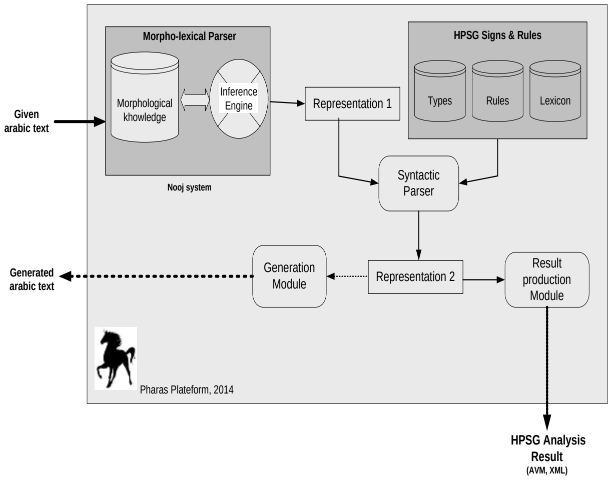


Figure 2: General architecture of Pharas platform

The "signs and rules" subsystem contains the HPSG signs, the type hierarchy and the rules to apply. It consists of three files: "Types", "Rules" and "Glossary".
In the " Type " file , we define the type hierarchy used to describe HPSG features. This file has an important role in HPSG since it is itself a set of constraints on feature structures.
The "Glossary " file contains all the lexical entries (verbs, nouns, adjectives, particles) already encountered or analysed.
Each entry is represented by an HPSG feature structure. Figure 3 gives an example of the Arabic verb " fahima " (he has understood). This structure contains all the features used to represent the inflected verbal form :
- Phon : briefly, it is the string corresponding to the inflected form,
- Class : the class of the verb,
- Voice : active or passive form,
- tense : past, present or imperative,
- mode : indicative, subjunctive, imperative,

- root : letters representing the root of the verb,
- Vform : Transitive or Intransitive,
- Person : first, second or third,
- number : singular, dual or plural,
- gender : masculine or feminine

Phon	فهم
Class	فعل
Voice	معلوم
Tense	الماضي
Mode	مبني على الفتح
Root	ف-ه-م
VForm	متعدي 2
Person	الغائب
Number	مفرد
Gender	مذكر

فهمم:= word & [PHON < ‘فهم’>, SS.LOC[CAT.Tete[MAJ فعل, VFORM 2, TENSE الماضي , ROOT ر-ح-م, MOD مبني على الفتح, CAT.Valence <>, CAT.S-ARG<>], CONT [Index[الفاعل:[PERSON الغائب, NUMB مفرد, GENDRE مذكر] , Restr <>]]]]

Figure 3: Example of features structure of verb "fahima" (to understand)

We remind that in HPSG, like the other full lexicalized formalisms , all of the constraints on the syntax and even semantics, are present in the lexical entries themselves. The File "Rules" contains all the syntactic rules to be applied to the standard Arabic language. The rules themselves are described by feature structures. Figure 4 gives an example of a syntax rule (head-complement rule), which determines whether complements follow or precede their head. This rule is used in the analysis of sentences such as فهم الولد الدرس ("the boy has understood the lesson").

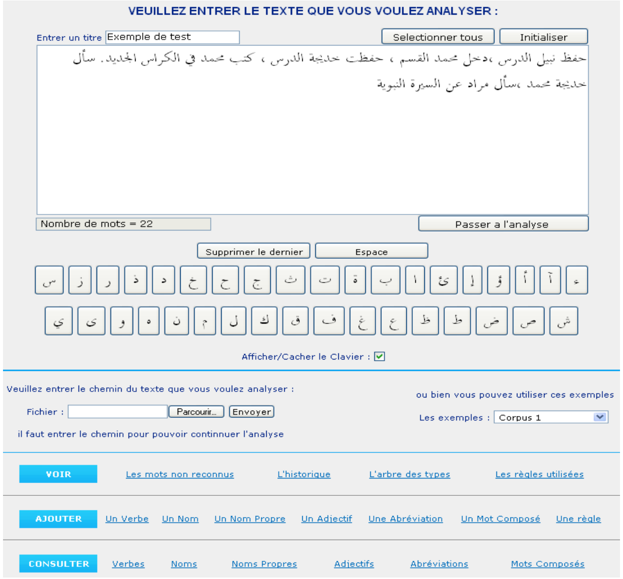
```
head-complement-rule-0 := phrase &
[ HEAD #0,
  SPR #a,
  COMPS #b,
  SEM #1,

  ARGS <  word & [ HEAD #0, SPR #a,
COMPS #b,  SEM #1] > ].
```

Figure 4 . Example of syntax rule : "head -complement "

3.3 Interface

Direct use of Pharas platform can be done via a set of interfaces. The main interface (Figure 5) gives the possibility for users to enter the text to be analysed from the keyboard or by loading it from a corpora. The user can then follow the different steps of analysis and find the rate of success and / or failure analysis of various items. The goal is to have the HPSG analysis, in XML format of the text given (see Figure 6) .



VEUILLEZ ENTRER LE TEXTE QUE VOUS VOULEZ ANALYSER :

Entrer un titre: Exemple de test

حفظ نبيل الدرس ،دخل محمد القسم ، حفظت خديجة الدرس ، كتب محمد في الكراس الجديد. سأل
خديجة محمد ،سأل مراد عن السيرة النبوية

Nombre de mots = 22

Afficher/Cacher le Clavier : ☑

Veuillez entrer le chemin du texte que vous voulez analyser :

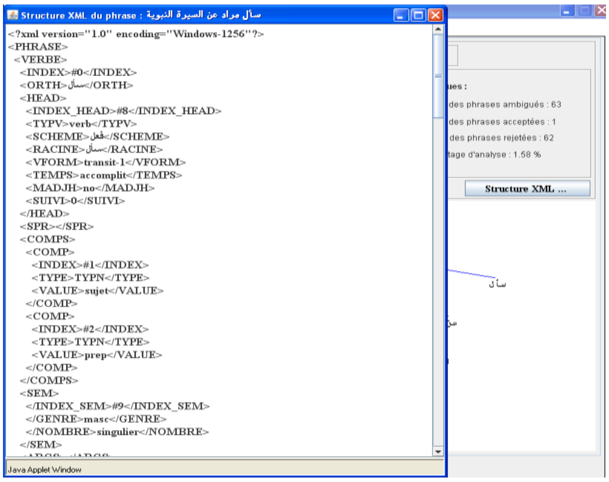
Fichier :

il faut entrer le chemin pour pouvoir continnuer l'analyse

ou bien vous pouvez utiliser ces exemples

Les exemples : Corpus 1

Figure 5. Main interface of Pharas platform



Structure XML du phrase : سأل مراد عن السيرة النبوية

```
<?xml version="1.0" encoding="Windows-1256"?>
<PHRASE>
 <VERBE>
  <INDEX>#0</INDEX>
  <ORTH>سأل</ORTH>
  <HEAD>
   <INDEX_HEAD>#8</INDEX_HEAD>
   <TYPV>verb</TYPV>
   <SCHEME>فعل</SCHEME>
   <RACINE>سأل</RACINE>
   <VFORM>transit-1</VFORM>
   <TEMPS>accomplit</TEMPS>
   <MADJH>no</MADJH>
   <SUIVI>0</SUIVI>
  </HEAD>
  <SPR></SPR>
  <COMPS>
   <COMP>
    <INDEX>#1</INDEX>
    <TYPE>TYPN</TYPE>
    <VALUE>sujet</VALUE>
   </COMP>
   <COMP>
    <INDEX>#2</INDEX>
    <TYPE>TYPN</TYPE>
    <VALUE>prep</VALUE>
   </COMP>
  </COMPS>
  <SEM>
   </INDEX_SEM>#9</INDEX_SEM>
   </GENRE>masc</GENRE>
   </NOMBRE>singulier</NOMBRE>
  </SEM>
```

ues :
des phrases ambigués : 63
des phrases acceptées : 1
des phrases rejetées : 62
tage d'analyse : 1.58 %

Figure 6. Example of analysis produced by Pharas platform

3.4 Morpho-Lexical Analysis Module

In an earlier version, this subsystem was developed using the "expert system" approach. The morpho-lexical knowledge was formalized in a set of rules and facts. Then inference engine was launched for any item to be analysed.

But since 2012, we have integrated a more efficient morpho-lexical module derivated from Nooj[1]. NooJ is a linguistic development environment, developed at Franche-Comté university in France, that includes large-coverage dictionaries and grammars, and parses corpora in real time. The Pharas partially uses the morpho-lexical step the services of the Nooj tool. Nooj offers a large set of linguistic resources (dictionaries, automata, etc ...) for several languages. When integrating the Nooj Tool, we confronted many challenges such as

[1] http://www.nooj4nlp.net/pages/nooj.html

adapting resources for processing Arabic language, ensuring proper interfaces between PHARAS platform and Nooj system , managing the interoperability problem, ... etc.

3.5 Syntactic Module

The syntactic analysis is based mainly on the process of unification. All AVM, for all lexical entries, produced and analysed in the previous step are parsed using rules.
The unification algorithm runs mainly in three stages:
- Indexing all the AVM of the sentence to be analysed.
- Application of existing syntactic rules, present in the system signs.
- Application of an incremental process of enrichment of feature structures obtained with the respect of the compatibility between structures.

The unification process stops after processing all the elements of the sentence and may give rise to the overall structure (if the text analysed contains no error) or fragments of structures (in case of unrecognized part of the text).

3.6 Result Generation Module

This module is responsible to return the analysis result to the user in one of the following formats: attribute-value matrix (AVM) or XML file.

3.7 Generation module

This module gives the reverse process of the analysis. The generation produces a text from its representation given in an AVM.

4. Resources and Linguistic Phenomena Covered

At the current state, the syntactic structure of standard Arabic language covered (which can be analysed with a success rate of over 90%) are: verbal sentences and nominal sentences.. The evaluation was made on a corpus composed of a set of standard Arabic texts (newspaper articles, parts of books).
To provide such coverage, many resources have been incorporated , including:
- A glossary of verbs composed of more than 2,000 verbs (triliteral and quadriliteral), distributed among the eight classes of the Arabic verbs. Each verb is stored with its HPSG feature structure .
- A glossary of common nouns of nearly 8,000 items including HPSG feature structures was constructed in a semi-automatically way from a corpus of newspaper articles
- A lexicon of adjectives nearly 1,000 items with their feature structures HPSG
- A glossary of proper names (people, places, etc ...) constructed in a semi-automatically way.
- All the particles used in Arabic (determiners, pronouns, prepositions, conjunctions, question, affirmation, negation, etc ...) ..

5. Conclusion

In this paper, we have presented a platform, based on HPSG formalism, for the standard Arabic language. The choice of HPSG is motivated by the effectiveness of this formalism in NLP researches.
The originality of the platform described is that it integrates, in the same environment, the whole chain of analysis modules (morpho-lexical, syntax and generation). Other existing tools would be limited to only one aspect.
Currently, the platform includes several resources (dictionaries and lexicons of thousands of nouns, verbs, adjectives, proper names, particles) constructed in manual or semi-manual way.
The syntactic subsystem covers , actually, verbal and nominal sentences of the standard Arabic language.
For further developments, several works can be undertaken, like:
- In terms of modelling: expanding the coverage of linguistic phenomena treated (passive sentences, interrogative, relative, coordination, ... etc. .) .
- In terms of implementation: developing APIs from platform which can be reused in NLP applications.

6. Acknowledgements

This project and publication were funded by the Algerian higher education and scientific Research Ministry.
The Authors would like to thank Max Silberztein, from the LASELDI laboratory at the Franche-Comté University for his approval for using Nooj tool in this project.
The authors would like to thank Pr Farid Meziane, from Salford University, for his valuable comments to improve the quality of the paper.

7. References

Anne Abeillé. (2007). Les grammaires d’unification, *Lavoisier Editions*.

Bahou Y., Hadrich Belguith L., Aloulou C., Ben Hamadou A. (2006). Adaptation et implémentation des grammaires HPSG pour l’analyse de textes arabes non voyellés., *Actes du 15e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle RFIA’2006*, Janvier 2006, Tours/France.

Bender Emily M. & Alex Lascarides (2013). On Modelling Scope of Inflectional Negation. In P. Hofmeister and E. Norcliffe (eds). The Core and the Periphery: Data-driven Perspectives on Syntax Inspired by Ivan A. Sag. Stanford: *CSLI Publications*. p.101-124.

Carl Pollard & Ivan A. Sag (1994). Head-driven Phrase Structure Grammar. Chicago: University of Chicago Press and Stanford: *CSLI Publications*.

Copestake Ann (2002). Implementing Typed Feature Structure Grammars , *CSLI Publications*, Stanford University, 2002.

Hann Michael (2011), Null Conjoncts and Bounds Pronouns in Arabic, *in Proceedings of HPSG 2011 Conference*, August 22-25 2011, University of

Washington, CSLI Publications.

Hann Michael (2012). Arabic Relativization Patterns: A Unified HPSG Analysis, *in Proceedings of HPSG 2012 Conference*, Chugnam National University of Daejon, South Korea, CSLI Publications, July 18-19 2012.

Hans C Bons & Ivan A.Sag (2012). Sign-Based Construction Grammar, *CSL Publications*, October 2012

Ivan A. Sag, Thomas Wasow and Emily M.Bender (2003) Syntactic Theory : a formal introduction, 2nd edition, *CSLI Publications*, ISBN 9781575.

Miyao Y. & Tsujii J. (2005). Probabilistic Disambiguation Models for Wide-Coverage HPSG Parsing , *In Proceedings of ACL-2005*, 2005, p. 83-90.

Mourad Loukam, Amar Balla & Mohamed Tayeb Laskri (2013). PHARAS : Une plate-forme d'analyse basée sur le formalisme HPSG pour l'Arabe standard : Développements récents et perspectives, *Revue RIST*, vol. 20, n° 2, p.20-31.

Ninomiya T., Matsuzaki T., Tsuruoka Y., Miyao and Y. Tsujii J (2006). Extremely Lexicalized Models for Accurate and Fast HPSG Parsing. *In Proceedings of EMNLP* ,2006.

Stephan Müller (2001). The Babel-System : a parser for an HPSG fragment of German, *Language Technology Lab*, Berlin University, 2001.

Rhythmic Features across Modern Standard Arabic and Arabic Dialects

Ghania Droua-Hamdani1, Yousef A. Alotaibi2, Sid-Ahmed Selouani3, Malika Boudraa4
1Speech Processing Laboratory, CRSTDLA
Algiers, Algeria
2College of Computer and Information Sciences, King Saud University
Riyadh 11543, Saudi Arabia
3LARIHS Laboratory, University of Moncton
Shippagan, Canada
4Signals and Speech Communication Laboratory, USTHB
Algiers, Algeria
E-mail: gh.droua@post.com, yaalotaibi@ksu.edu.sa, sid-ahmed.selouani@umoncton.ca, mk.boudraa@yahoo.fr

Abstract

This paper describes the speech timing of Modern Standard Arabic (MSA) using rhythmic features. Two main approaches are used to compute rhythm metrics: Interval Measures and Pairwise variability indices. The first approach consists of a comparison of MSA rhythm metrics computed from the ALGerian Arabic Speech Database with those resulting from using the West Point Arabic Speech corpus. The second approach compares these results with the rhythm measurements of six Arabic dialects. Many unexpected results are observed concerning rhythm variation within MSA and the dialects according to speakers' localities.

Keywords: rhythm metrics; Modern Standard Arabic; Arabic dialects

1. Introduction

Recent studies have developed a number of metrics to quantify rhythm in languages. These metrics have contributed to the discovery of new insights into how speech timing functions both across and within languages. Basically, rhythm metrics are computed from the acoustic durations of vocalic and consonantal intervals in continuous speech. Examples of experimentation addressed on languages such as English, French, Italian, Spanish, etc. (Arvaniti, 2009; Nolan & Asu, 2009; Giordano & D'Anna, 2008; White & Mattys, 2007-a). Regarding Arabic language, a comparison between Modern Standard Arabic (MSA) and other languages was conducted in to works (Droua-Hamdani, et al, 2010; Selouani, Alotaibi & Pan, 2012). In addition, rhythm of Arabic dialects was studied by Hamdi et al. (2004).

The paper deals with both MSA language and Arabic dialects. In terms of linguistic structure, there are phonetic, morphological, lexical, and syntactic differences between these both categories of Arabic.

In this paper, two investigations are conducted. The first, an experiment, concerns the comparison of MSA rhythm metrics computed from speech sentences of the ALGerian Arabic Speech Database (ALGASD corpus) pronounced by Arabic western speakers (Algerians) (Droua-Hamdani, Selouani & Boudraa, 2010) with those calculated from the West Point Arabic Speech corpus (LDC). The second investigation involves a comparison of the MSA rhythm metrics obtained by both the ALGASD and West Point corpora with the dialect rhythm metrics found by Ghazali et al. (2002) and Hamdi et al. (2004). In the study, six Arabic dialects were chosen from different geographic areas: Moroccan, Algerian, Tunisian, Egyptian, Lebanese, and Jordanian. The speech data were taken from free translations of the fable The North Wind and the Sun into each native speaker's dialect.

The purpose of this investigation is to highlight the variations in rhythm metrics across the Arabic language (MSA and dialects) according to the speakers' localities.

The paper is organized as follows: Section 2 gives a description of the main approaches used to compute the rhythm metrics; Section 3 describes the ALGASD and West Point corpora used to compute the MSA rhythm metrics; Section 4 presents the experimental results. Conclusions and future perspectives are given in the last section.

2. Rhythm metrics

Based on the acoustic durations of vocalic and consonantal intervals in continuous speech, there are generally two main approaches used to compute rhythm metrics: Interval Measures (IM) (Ramus, Nespor & Mehler; 1999) and Pairwise variability indices (PVI) (Grabe & Low; 2002).

The IM approach involves computing three separate measures from the segmentation of speech signals into vocalic and consonantal units. IM metrics and their normalization (Dellwo, 2006) are as follows:

- ΔV: standard deviation of vocalic intervals
- ΔC: standard deviation of consonantal intervals
- %V: percentage of utterance duration composed of vocalic intervals
- VarcoV: standard deviation of vocalic intervals divided by the mean vocalic duration
- VarcoC: standard deviation of consonantal intervals divided by the mean consonantal duration

The PVI approach, in contrast, aims to express the level of variability in successive vocalic and intervocalic intervals. The nPVI measure computes the normalized differences of subsequent vocalic durations, and the rPVI score calculates the successive intervocalic intervals (consonants). The nPVI and rPVI are defined, respectively, by:

$$nPVI = 100 \times \frac{\left(\sum_{k=1}^{m-1} \left| \frac{(d_k - d_{k+1})}{(d_k + d_{k+1})/2} \right| \right)}{(m-1)} \qquad (1)$$

$$rPVI = \frac{\left(\sum_{k=1}^{m-1} \left| d_k - d_{k+1} \right| \right)}{(m-1)} \qquad (2)$$

where m is the number of intervals, and d is the duration of the kth interval.

3. Methodology

3.1 Data collection

The ALGASD was originally designed to train and test automatic speech recognition engines. However, it is now also used in several areas of research, such as a study based on rhythm metrics to show the typology of Algerian MSA among a set of languages (Droua-Hamdani, et al, 2010).

ALGASD reflects the main variations of pronunciation in MSA due to the regional and social differences of speakers in Algeria, including gender, age, education level, and MSA mastery level. The database consists of 1,080 recordings collected from 300 speakers recruited from 11 distinct regions (Droua-Hamdani, Selouani & Boudraa, 2010). All sentences were recorded on an individual basis in a quiet environment at a sampling rate of 16 kHz. Speakers were instructed to read at a comfortable rate and in a normal voice.

The West Point corpus contains MSA speech data collected and processed by members of the Department of Foreign Languages at the United States Military Academy at West Point and the Center for Technology-Enhanced Language Learning (CTELL) (LDC). The original purpose of this corpus was to train acoustic models for automatic speech recognition using Arabic (MSA) speech files recorded by both native and non-native speakers (110 speakers). The place of origin of the native speakers is not noted, but their accents seem to be from the Middle Eastern. Recordings were captured at a sampling rate of 16 bit at 22,05 kHz.

3.2 Measurements

For the current experiment, we used the data from 24 ALGASD speakers (12 female and 12 male). The text material includes 22 sentences. As regard to West point corpus, speech material from 15 Arabic speakers (10 female and 5 male) reading five sentences was used.

All vowels and consonants were segmented manually by inspection of the respective speech waveforms and wideband spectrograms. Vowel and consonant durations were extracted using a customized script on the boundary label files.

The 7 rhythm metrics examined across all sentences are: three interval measures (%V, ΔV, and ΔC), two time-normalized indices (VarcoV and VarcoC), and two pairwise variability indices (nPVI-V and rPVI-C).

4. Results

4.1 ALGASD vs. West Point rhythm metrics

The first experiment compared MSA rhythm metrics calculated from sentences spoken by Western speakers (North Africa, from ALGASD) vs. Eastern ones (West Point). Table 1 reports mean values of each of the seven rhythm metrics as applied to both groups. The vocalic interval measure and time-normalized intervals (ΔV, VarcoV and VarcoC) resulting from the Eastern group are higher than the rhythm metric scores from the Western participants. Likewise, the vocalic and consonantal pairwise variability indices (nPVI-V and rPVI-C) and speech rate are higher than in the ALGASD samples. The %V value for West Point is lower than the ALGASD measure. For both samples, ΔC is very similar.

Table 1 shows results of one-way analyses of variance (ANOVAs) that test for differences in rhythm metrics between both corpora. Five of the seven metrics – the two interval measure metrics, both pairwise variability indices and vocalic time-normalized metric– are sensitive to speakers' geographical origins. Consonant proportion (ΔC) and time-normalized consonant metrics (VarcoC) show no significant effects.

	ALGASD	West Point	F	P
%V	46.2	42.4	F(1,165)=11.46	P<.000
ΔV	39.19	49.42	F(1,165)=16.64	P<.000
ΔC	53.90	53.29	F(1,165)=0.39	P=.84
VarcoV	37.86	65.14	F(1,165)=21.42	P<.000
VarcoC	48.57	50.81	F(1,165)=2.184	P=.141
rPVI-C	56.47	74.84	F(1,165)=28.97	P<.000
nPVI-V	37.56	56.22	F(1,165)=57.81	P<.000

Table 1: Mean values of rhythm metrics (ms) and results of one-way ANOVAs testing effect of speakers' origins

There is a significant difference in the duration of vowels in MSA by Algerian (Western) speakers compared with Eastern speakers. The percentage of vocalic durations in syllable structure produced by in the West Point corpus is lower than for the Western speakers. When comparing both %V values (Eastern/Western) to other languages, it was found that the %V value computed for West Point approached the vocalic percentage measures of stress-timed languages (e.g. Dutch: 41%; English: 38%). The ALGASD %V measure is close to syllable-timed languages such as French and Spanish (45% and 48% respectively) (White & Mattys, 2007-b). These unexpected results show that the pronunciation of vowels by Eastern speakers decreases in terms of vocalic duration when compared with the ALGASD Western speakers. This vocalic reduction makes syllable structures appear more complex in the West Point MSA than for the ALGASD MSA. The vocalic rhythm metrics presented could be used as distinctive parameters within MSA to categorize speakers according to their geographic locations (Eastern vs. Western).

To reveal possible differences in rhythm patterns of speakers according to their gender, all seven metrics were calculated for both corpora (Figure 1). The vocalic metrics (ΔV, VarcoV and nPVI-V) and the consonantal pairwise variability index (rPVI-C) of ALGASD male speakers show higher scores than ALGASD female speakers. %V values of both classes of speakers are closer to each other. The consonantal time-normalized interval (VarcoC) and the consonantal interval (ΔC) of ALGASD's males are lower than their female counterparts. For West Point speakers, the results show that four metrics (%V, VarcoV, VarcoC and nPVI-V) calculated for males are higher than the values computed for females from that corpus. The ΔV and rPVI-C of West Point males are lower than West Point females' rhythm metric values.

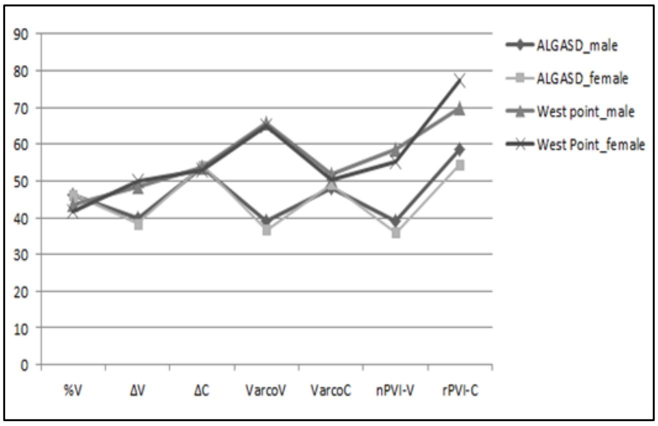


Figure 1: Effect of gender on rhythm metrics for ALGASD and West Point speakers

Two series of ANOVAs were performed to test the effect of gender on all seven rhythm metrics of the corpora. No significant effect for gender was found.

4.2 MSA vs. Arabic dialects rhythm metrics

The second investigation compares the MSA rhythm metrics and findings of [5, 10] for Arabic dialects. Six dialects were studied: 3 Arabic western dialects (Algerian, Moroccan, and Tunisian) vs. 3 Arabic eastern dialects (Egyptian, Lebanese, and Jordanian).

Table 2 shows the values of %V, ΔC, nPVI-V, and rPVI-C for each dialect in contrast with the MSA rhythm measures from both the ALGASD and West Point corpora. As can be seen from Table 2, %V of ALGASD MSA is higher than all vocalic percentage measures finding for all dialects. However, %V value of West Point is close to Jordanian and Lebanese dialects' scores. ΔC metrics of both MSA corpora are closer to the Middle Eastern dialects (Egyptian, Lebanese, and Jordanian) than the western ones, such as the Algerian and Moroccan dialects. nPVI-V metrics show that the MSA language presents higher values than Eastern and Western Arabic dialects. rPVI-C of West Point is close to Western dialects (Algerian and Moroccan) than to the Middle Eastern dialects (Egyptian, Lebanese, and Jordanian). So, these results conclude that, for example, the Algerian dialect presents higher consonantal proportions (ΔC and rPVI-C) and lower vocalic rhythm values (%V, and nPVI-V) than ALGASD MSA which is recorded from Algerian speakers. These findings further suggest that in comparison to the MSA, when using Arabic dialects, western speakers reduce the vocalic intervals that help to sustain syllable structures. This vocalic reduction leads the syllables to appear more consonantal in dialects. As regard to the IM measures, the vocalic proportions produced by eastern speakers (Lebanese, and Jordanian) are relatively kept between the MSA of West Point and Arabic dialects. So, the Arabic pronunciation of vowels and consonants in terms of rhythm metrics are similar for these both categories of language for eastern speakers.

	Language	%V	ΔC	nPVI-V	rPVI-C
MSA	ALGASD	46.20	53.90	53.21	53.46
	West Point	42.4	53.29	56.22	74.84
Dialects	Moroccan	33.14	72.68	46.50	79.89
	Algerian	33.10	68.10	46.08	78.73
	Tunisian	35.42	56.85	44.41	63.74
	Egyptian	37.41	53.67	45.53	57.37
	Lebanese	41.63	54.55	47.05	61.02

Table 2: Comparison between MSA and dialects rhythm metrics

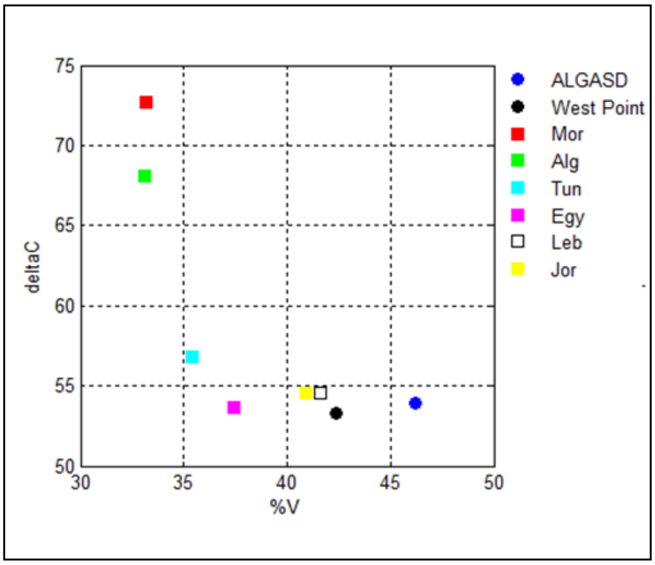
In the field of rhythm categorization, two IM measures are widely used to classify languages: ΔC and %V. With respect to the PVI approach, the measures of nPVI-V and rPVI-C are also used. Figure 2 and Figure. 3 show the plane projections of these IM measures (ΔC and %V; nPVI-V and rPVI-C) for all Arabic dialects and MSA. Both figures show the location of West Point MSA and ALGASD MSA among all dialects: Algerian (Alg), Moroccan (Mar), Tunisian (Tun), Egyptian (Egy), Lebanese (Leb), and Jordanian (Jor).

Results show that the location of West Point MSA in the plane projection (ΔC, %V) is near the eastern dialects (Figure 2). However, ALGASD MSA is far from western dialects. Indeed, it presents a proportion of ΔC and similar to that of the eastern dialects.

Regarding Figure 3, the PVI approach shows that the deviation between ALGASD and West Point MSA are far from all dialects. They are in the left and right top of the chart.

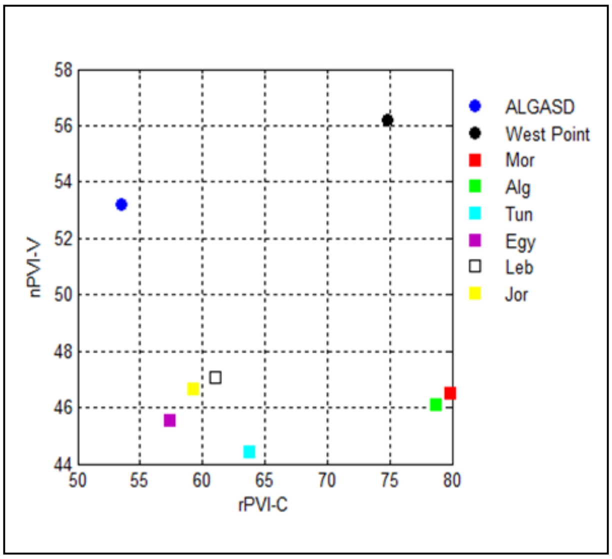
5. Conclusion

The paper deals with the rhythmic features across MSA language and various Arabic dialects. A comparison between rhythm metrics computed from two MSA corpora (ALGASD vs. West Point) was conducted. The results showed that speakers of ALGASD, all of whom are Algerians (i.e., western speakers), present several differences in rhythm scores compared with West Point native (i.e., eastern speakers). These rhythm variations are noted using both the IM and PVI approaches for both vowel and consonant scores. According to these results, it thus seems that eastern speakers produce longer vowels durations than western speakers, despite using the same language (MSA).



Alg: Algerian, Mor: Moroccan, Tun: Tunisian, Egy: Egyptian, Leb: Lebanese, Jordanian: Jor.

Figure 2: Comparison of MSA and Arabic dialects in (ΔC, %V) plane



Alg: Algerian, Mor: Moroccan, Tun: Tunisian, Egy: Egyptian, Leb: Lebanese, Jordanian: Jor.

Figure 3: Comparison of MSA and Arabic dialects in (nPVI-V, rPVI-C)

In addition, the rhythm metrics of both corpora were compared in a second experiment. The results revealed that MSA provides rhythmic variations when compared with the rhythm metrics of all dialects (eastern and western).

The results of both of these experiments show that rhythmic parameters can be used as discriminant parameters in Arabic. Indeed, as shown by the current results, rhythm measures are relevant features for the classification of MSA speakers belonging to different Arabic regions (e.g., eastern vs. western). The observed differences in MSA pronunciation occurred across vowel and consonant durations. Finally, the current rhythm metric results also show that such metrics are pertinent features that can assist in distinguishing between MSA and the many Arabic dialects. The results are obtained for speakers that are adults but it is interesting to pursue the work to show if the Arabic variations and the prosodic tagging are similar to these findings when speakers are children.

These primary results will be added to other prosodic parameters to further improve the ability to distinguish between these languages. These features will be used in the conception of an automatic system of Arabic classification.

6. Acknowledgements

This work is supported by the NPST program at King Saud University, Project Number 10INF1325-02.

7. References

Arvaniti, A. (2009). Rhythm timing and the timing of rhythm. *Phonetica*, 66, pp. 46--63.

Nolan, F., Asu, E. (2009). The pairwise variability index and coexisting rhythms in language. *Phonetica,* 66, pp. 64--77.

Giordano, R., D'Anna, L. (2008). A comparison of rhythm metrics in different speaking styles and in fifteen regional varieties of Italian. In Speech Prosody.

White, L., Mattys, S.L. (2007). Rhythmic typology and variation in first and second languages. *Segmental and Prosodic issues in Romance Phonology*. pp. 237--257.

Droua-Hamdani, G., Selouani, S.A, Boudraa, M., and Cichocki, W. (2010). Algerian Arabic rhythm classification. ISCA International Speech Communication Association, in Proceedings of the third ISCA Tutorial and Research Workshop Experimental Linguistics, ExLing2010. pp. 37--41

Selouani, S-A., Alotaibi, Y.A., and Pan, L. (2012). Comparing Arabic rhythm metrics among other languages. In: Audio, Language and Image Processing (ICALIP). pp. 287—291.

Hamdi, R., Barkat-Defradas, M., Ferragne, E., and Pellegrino, F. (2004). Speech timing and rhythmic structure in Arabic dialects: a comparison of two approaches. Interspeech04,

Droua-Hamdani, G., Selouani, S-A., and Boudraa, M. (2010). Algerian Arabic Speech Database (ALGASD): corpus design and automatic speech recognition application. *Arabian Journal for Science and Engineering*. 35, Number 2C, 157--166,

http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S02

Ghazali, S. Hamdi, R. and Barkat-Defradas, M. (2002). Speech rhythm variation in Arabic dialects. Speech Prosody 2002. .

Ramus, F., Nespor, M. and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73, 265—292.

Grabe, E. Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. Papers in Laboratory Phonology 7, pp. 515—546.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for deltaC. Language and Language Processing. Paper presented at the 38th Linguistic Colloquium, 231241 (Peter Lang, Frankfurt, edited by P. Karnowski and I. Szigeti).

White L., Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, 35(4), pp. 501—522.