YADAC: Yet another Dialectal Arabic Corpus

Rania Al-Sabbagh, Roxana Girju
Department of Linguistics, University of Illinois at Urbana-Champaign
61801 Urbana, IL, USA
alsabba1@illinois.edu, girju@illinois.edu

Abstract

This paper presents the first phase of building YADAC – a multi-genre Dialectal Arabic (DA) corpus – that is compiled using Web data from microblogs (i.e. Twitter), blogs/forums and online knowledge market services in which both questions and answers are user-generated. In addition to introducing two new genres to the current efforts of building DA corpora (i.e. microblogs and question-answer pairs extracted from online knowledge market services), the paper highlights and tackles several new issues related to building DA corpora that have not been handled in previous studies: function-based Web harvesting and dialect identification, vowel-based spelling variation, linguistic hypercorrection and its effect on spelling variation, unsupervised Part-of-Speech (POS) tagging and base phrase chunking for DA. Although the algorithms for both POS tagging and base-phrase chunking are still under development, the results are promising.

Keywords: Dialectal Arabic, Dialect Identification, POS tagging

1. Introduction

Dialectal Arabic (DA) refers to a large number of Arabic dialects that speakers in the Arabic-speaking world acquire as their native languages. Despite sharing a considerable number of semantic, syntactic, morphological and lexical features with one another and with Modern Standard Arabic (MSA) variety, Arabic dialects do substantially differ in almost all language subsystems (i.e. semantics, syntax, morphology and phonetics … etc.). NLP interest in DA has increased recently given that it is the language variety of Arabic mostly used in chats, microblogs, blogs, forums, informal email, many recent TV shows and newspapers, which are themselves the target for NLP tasks and applications such as sentiment analysis and opinion extraction.

Current available corpora for DA – namely the LDC CallHome and CallFriend series (Canavan and Zipperlen 1996, Canavan et al. 1997) and ELDA – are small in size and focus on spoken rather than written DA. That is why there are many ongoing research projects trying to build DA corpora that represent the language in its written variety, especially as used on the Web. Different resources are used such as blogs and forums (Diab et al. 2010); online readers' commentaries on newspaper posts (Zaidan and Callison-Burch 2011) and also Web resources combined with more traditional resources like books and newspaper articles as in McNeil and Faiza (2011). Moreover, most of current efforts aim at annotating the compiled corpora for basic linguistic information such as: Part-of-Speech (POS) tags, degree of dialectness, and sentence boundaries among others.

YADAC is meant as a multi-dialectal and multi-genre corpus for DA. Although this paper presents the first phase of the corpus compilation and analysis, which focuses on Egyptian Arabic (EA), future work involves both Levantine and Gulf Arabic. The first contribution brought by our corpus – in addition to being currently larger than most of the complied corpora as discussed in section (8) – is brining two new genres to DA corpus complication efforts, namely microblogs (i.e. Twitter) and online knowledge market services. With their conversation-like direct interaction, Twitter and QA posts make our corpus not only usable for information extraction (i.e. POS tags, base-phrases, collocations, parsing … etc.), but also for more computer-mediated human interaction studies, that work on how Web users communicate to express opinions, show sentiment and take sides in arguments.

The second contribution of YADAC is using function words to build the Web harvesting search queries and also to build a threshold model for dialect identification. The main assumption is that using lexical content words might not be reliable enough given the overlap between DA and Modern Standard Arabic (MSA) that is experimentally proved (Duh and Kirchhoff 2005).

Highlighting and tackling spelling variation in DA corpora as caused by vowel-based variations and linguistic hypercorrection are the third contributions of YADAC. Finally, we offer basic linguistic analyses represented by POS tagging and base-phrase chunking. Despite the many performance improvements that are still to be made in future phases of the corpus, the preliminary results are promising.

The rest of this paper is divided into 9 sections. The first section is about the Web harvesting process and the search queries used for that purpose. The second section discusses the function-word threshold model used for filtering corpus items of zero- or low-dialectal content. The third section is about spelling variation in DA and how it is handled given information extraction as our testing platform. It also handles spelling hypercorrection which is a less commonly tackled problem in DA spelling variation. The fourth section deals with the POS tagging preliminary results and ways to improve performance and the same thing is done for base-phrase chunking in section five. Samples of actual searches through the corpus and corpus descriptive statistics are given in section six before a brief summary of related work and concluding remarks in sections, 7 and 8, respectively.

2. Web Harvesting DA

Three venues are used for corpus compilation: Twitter-API-based search engines, online knowledge market services and blog-based search engines. For each one of these, two different search engines are used to overcome the upper-bound limit of the returned search results that each search engine sets per query.

Generic queries, each of which consists of a minimum of three function words, are automatically created by permuting the entries of a 1,527 EA-exclusive function words list. Using function words is meant to create topic-independent search queries and thus broaden the search scope and harvest more data. Out of the created permutations, 15M search queries are randomly selected and used to crawl the Web over a period of 7 months – May 2011 to November 2011.

An example of the used search queries is: " ** مش ** إزاي عشان" /*<zy* ** *m$* ** *E$An*/ (how ** not ** because). The asterisk stands for multiple words in all the used search engines. There are two hypotheses for using this type of search queries: first***, EA-exclusive function words are generic words that users are to use in their EA posts regardless of the topic***; second, ***multi-word search queries are likely to return search hits with bigger chunks of written material***; unlike mono-word search queries that can return posts – especially blog posts – in which the word is mentioned in a video, song or photo title.

The crawling output is then cleaned from HTML markup, noisy results (i.e. spam, advertisements, video and audio results, and broken links). Another cleaning step is removing any overlapping search results across the different search engines. An output of 11M words resulted from this Web harvesting process. Each harvested item (i.e. tweet, Question-Answer (QA) pair from the online knowledge market services, or blog post) has to be filtered based on its degree of dialectness.

3. A Threshold Module for Dialectness

Although the search-query sets are designed to be dialect exclusive, Arabic varieties overlap is almost inevitable. Not only does MSA overlap with Arabic dialects, but also Arabic dialects overlap across one another. The size of the overlap can range from single words to complete phrases or clauses. For the purpose of our corpus, we focus only on written dialectal identification, unlike most work on dialect identification that focuses on spoken corpora and thus relies on prosodic phonological features **(**Alorifi 2008 and Biadsy et al. 2009 among others).

Both Diab et al. (2010) and Zaidan and Callison-Burch (2011) build their dialect identification models using content words. Using the MSA morphological analyzer – Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter 2004), Diab et al. (2010) make the hypothesis that if BAMA is unable to generate a morphological analysis for an input word, it is then a DA word. According to this hypothesis, 19%, 13.5$, 8% and 26% of the unigram word types in Egyptian, Iraqi, Levantine and Moroccan blog posts are assumed DA-exclusive, respectively, in their COLABA corpus. Out of these claimed DA-exclusive words, 35% are dialectal words and 30% are named entities. Moreover, 50% and 25% of the least frequent bigrams and trigrams, respectively, involve at least one dialectal word. The percentages of named entities in bigrams and trigrams are 19% and 43%, respectively. Zaidan and Callison-Burch (2011) use a trigram-based model for dialectal identification built according to online readers' commentaries on newspapers posts that are manually evaluated by native speakers as being highly dialectal. The model achieves a precision rate of 71.2% and a recall rate of 77.6%.

MSA and EA, on one hand, and EA and other Arabic dialects, on the other hand, share a considerable part of their lexical repositories. This is proved by Duh and Kirchhoff (2005) using BAMA that gave analyzes to 62.8% of their EA corpora and 71.8% the Levantine Arabic corpora. Words analyzed by BAMA can be divided into three categories:

(1) words that have the same phonetic, lexical, morphological and syntactic features in both MSA and DA like الجيش /*Aljy$*/ (the army);

(2) words that have the same lexical, morphological and syntactic features but different pronunciation like قال /*qAl*/ (said); which is pronounced with as a voiceless uvular plosive in MSA and with a glottal stop in EA despite being written with the voiceless uvular plosive ق /*q*/ in both dialects;

(3) words that have the same phonetic features, but different lexical meanings, grammatical categories and morphological features as in بيئة /*by}p*/ which is a noun meaning *environment* in MSA but an adjective in EA meaning *vulgar*.

That is why our dialectal identification model relies on EA-exclusive function words and affixes to set a coarse threshold to filter out corpus items (i.e. tweets, QA pairs and blog posts) of low- or zero-dialectal content. The hypothesis is that ***dialect-exclusive function words and affixes; surrounding content words, are better cues for whether a content word is being used in its dialectal or standard meaning in case it belongs to any of the aforementioned three categories***. A set of 1,000 randomly selected items – see appendix (A) for a sample – is manually evaluated by two EA native speakers on a scale of 1 to 3, with 1 being mostly MSA and 3 being mostly DA. With an inter-annotator Kappa score of 0.8, 730 items are evaluated as 3. Diagram (1) show the precision and recall rates for function words alone and for using them with the highly frequent EA-exclusive prefix بـ /*b*/ (aspectual progressive prefix as in بيكتب /*byktb*/ - he's writing). Although the relatively low recall rates lead to losing many corpus items, our module guarantees high dialectal content. The corpus size after dialect identification is approximately 6M words.

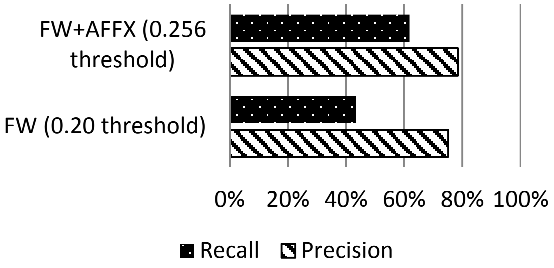

Diagram 1: Precision and recall rates for the threshold model of dialect identification

4. Spelling Variations

Spelling variation due to the lack of a conventional standard tradition of writing has always been claimed a problem in DA corpora. We claim here that most of DA spelling variations can be traced back to phonetic and phonological differences between MSA and DA. . In dealing with these differences in writing, some speakers prefer to retain the MSA spelling despite using the DA pronunciation to show their decent educational background; while others simply write the way they speak. For example, in EA the voiceless uvular plosive *qaf* ق /*q*/ is by default debucclized into the glottal stop أ />/. Given that, the MSA قوي /*qwy*/ (very) is always pronounced with a glottal stop initial as in أوي />*qy*/ (very). Although all EA speakers pronounce it with the initial glottal stop, some write it with a MSA-based spelling *qaf* (312 occurrences in our corpus) and others with the DA-based glottal stop (1,557 occurrences).

Habash et al. (2011) and Dasigi and Diab (2011) give an extensive overview of the spelling variation phenomena in DA, with Dasigi and Diab (2011) developing a system to automatically conflate variation and thus overcome data sparseness issues. Examining these overviews, we can classify consonant-based spelling variations in 5 classes illustrated in table (1). Not all differences between MSA and DA are relevant to all Arabic dialects, however. For instance, Habash et al. (2011) worked on cases like the difference between the MSA كحك and the DA كعك which is not relevant for YADAC at this phase. Since كعك is almost never used in spoken or in written EA, this spelling variation does not almost exist in YADAC. Some words have been adapted from MSA with the MSA origin completely neglected in a given Arabic dialect.

Vowel-based spelling variation is usually neglected although orthographically-represented vowel lengthening is frequent in EA corpora. Typically, the short vowels */a/*, */i/* and */u/* are not represented in written corpora. This is because they are orthographically represented using the diacritics – (َ), (ِ) and (ُ), respectively, and diacritics are rarely used in written MSA corpora and almost never in DA written corpora. However, the long vowels */a:/*, */i:/* and */u:/* are orthographically represented by the *alef* ا */A/*, and the glides *ya'* ي */y/* and *waw* و */w/*, respectively. Vowel-based spelling variation occurs due to missing short vowels for long vowels. For instance, the short vowel /*i*/ in انتِ /*Anti*/ (you; singular feminine) is frequently misspelled as *ya'* انتي */Anty/* - (548 vs. 1,295 occurrences, respectively). Similarly, the long vowel *ya'* ي */i:/* is frequently inserted after the first letter in فلسطين */flsTyn/* (Palestine) instead of the correct short vowel */i/*; 2409 occurrences compared to 1,837 occurrences for the correct short-vowel version.

Linguistic hypercorrection refers to pronunciation or grammatical constructions produced by mistaken analogy with the standard usage out of a desire to be correct. EA speakers who are aware of the aforementioned DA spelling variations try to adhere to MSA writing conventions to add social status to their written production. As a result, they sometimes over correct or over-standardize many such variations. For instance, although قوله /*qwlh*/ (tell him) is correctly pronounced and written in both MSA and EA with a long vowel after the first letter, speakers can mistakenly hyper-correct it by deleting the long vowel and its glide representative و */w/*. As a result, the correct قوله */qwlh/* (tell him) occurs 2,207 times in YADAC and the erroneous hypercorrection قله */qlh/* (tell him) 1,987 times.

There are other sources of DA spelling variation. Word lengthening, for example, as in كتيييير */ktyyyyr/* (a lot) or عاااالي */EAAAAly/* (high) is very common. Borrowed proper nouns are another source as in *Facebook* that can be written as one word فيسبوك */fysbwk/* (2944 occurrences) or two words فيس بوك */fys bwk/* (1033 occurrences). There are also unpredictable – or phonologically unjustifiable – spelling-variation cases like *too*; which can be written as برضه */brDh/*, برضك */brDk/* or برضو */brDw/*. This is an EA-exclusive word and thus speakers have no other choice but writing it the way they pronounce it; which has three variations.

For the purposes of this phase of YADAC and for the current application for which it is used (i.e. information extraction), each search query is mapped to all its possible spelling variations taking into consideration consonant- and vowel-based variations, the effect of hypercorrection, word lengthening and also using a list of 138 cases of unpredictable spelling variations. Spelling mapping follows a corpus-based approach according to which each search query is tested for the aforementioned spelling variations and only spelling variations found in the corpus are considered as valid and results from them are also returned alongside the results of the original search query. Therefore, submitting اكتر */Aktr/* (more) – 280 occurrences – as a search query results in finding results also for أكتر */>ktr/* (166 occurrences), اكثر */Akvr/* (41 occurrences) and أكثر */>kvr/* (54 occurrences).

5. POS Tagging

We use our hybrid-approach tagger to simultaneously tokenize and POS tag YADAC (Al-Sabbagh and Girju

	Change	Examples			
		MSA Form	**Corpus Frequency**	**EA Form**	**Corpus Frequency**
Fricatives	**Voiceless dental fricative → Voiceless alveolar plosive**	ثمن */vmn/* (price)	673	تمن */tmn/* (price)	3,150
	Voiced dental fricative → Voiced alveolar sibilant	ذمة _/*mp/_ (protection)	147	زمة */zmp/* (protection)	230
	Voiced dental fricative → Voiced alveolar plosive	كذب _/k*b/_ (lying)	1,074	كدب */kdb/* (lying)	33
Plosives	**Voiceless Uvular Plosive → glottal stop**	قوي */qwy/* (very)	422	أوي */>wy/* (very)	1,775
Pharyngealized	**Pharyngealized voiced alveolar sibilant → Pharyngealized voiced dental fricative**	ضابط */DAbT/* (officer)	1,427	ظابط */ZAbT/* (officer)	3,342
	Pharyngealized voiced dental fricative → Pharyngealized voiced alveolar sibilant	ظلمة */Zlmp/* (darkness)	19	ضلمة */Dlmp/* (darkness)	16
	Pharyngealized voiced alveolar sibilant → voiced alveolar plosive	ضحك */DHk/* (laughter)	246	دحك /dHk/ (laughter)	100
Glottal Stop (Hamza)	**Hamza deletion with almost all short and long vowels**	أمريكا */>mrykA/* (America)	8,983	امريكا */AmrykA/* (America)	2,206
		إجازة */<jAzp/* (vacation)	404	اجازة */AjAzp/* (vacation)	361
		جائزة */jA}zp/* (prize)	79	جايزة */jAyzp/* (prize)	8
		رئيس */r}ys/* (boss)	4,517	ريس */rys/* (boss)	351
		مسؤول */ms&wl/* (responsible)	784	مسؤل */ms&l/* (responsible)	620
		سماء */smA'/* (sky)	16	سما */smA/* (sky)	230
Dotted Consonants	**Word-final voiceless alveolar plosive *ta' marbouta* (feminine marker) → voiceless glottal fricative**	ناشطة */nA$Tp/* (activist)	347	ناشطه */nA$Th/* (activist)	453
	Word-final palatal approximant → near-open front unrounded vowel *(alef maqsura)*	حقي */Hqy/* (my right)	4,193	حقى */HqY/* (my right)	763

Table 1: Phonologically-based Spelling Variations across MSA and EA Consonants

2011). Although the tagger is still under development, preliminary results are quite promising. The tagger is built at three phases that aim at minimizing manual annotations for the training corpus to maximize its size. The first phase relies on using large raw corpora and a Finite-State Transducer (FST) module to simultaneously tokenize and POS tag the raw corpus using word-level inflectional morphology information represented by affixes and clitics.

The used set of affixes include tense-based affixes like the present-tense prefix يـ */y-/* in يكتب */yktb/* (he write), aspect-based affixes like the aspectual progressive prefix بـ */b-/* in بيكتب */byktb/* (he's writing), number-based affixes like the plural suffix ين */-yn/* in مصريين */mSryyn/* (Egyptians) and gender-based affixes like the feminine suffixes ـة */-p/* in حلوة */Hlwp/* (beautiful; singular feminine adjective). This affix sets enables semantic-feature labeling, in addition to tokenization and POS tagging. The set of clitics includes object pronouns, possessive pronouns and negative circumfixes among other clitics. Although the conjunction و */w/* (and) is not a clitic, it is conventionally written as one – i.e. attached to the beginning of the words.

The FST module is divided into two sub-modules: an analyser and a generator. The analyser module starts chopping off one affix or clitic at a time, bi-directionally (i.e. right-to-left and then left-to-right) while checking the validity of every analysis output against the corpus to prevent over-analysis. When reaching the shortest possible wordform (i.e. further chopping leads to invalid wordforms according to the corpus; the shortest valid wordform is assumed affix- and clitic-free), the generator module is activated. The generator reverses the analysis process and adds one affix or clitic at a time bi-directionally while blocking affixes and clitics used by the analyser for the same word to prevent duplicates. The output of each generation is validated using the corpus to prevent over-generation.

With manually-annotated gold standard set of 3,000 words – 1,000 words for each genre, the FST module and its two sub-modules show consistent performance rates across the three genres in our corpus as in tables (1a-c).

		Recall	**Precision**	**F-Measure**
Twitter	**TOK**	0.95	0.94	0.945
	SF	0.639	0.775	0.701
	POS	0.901	0.897	0.899
	ALL	0.869	0.827	0.847

		Recall	**Precision**	**F-Measure**
QA Pairs	**TOK**	0.982	0.954	0.967
	SF	0.65	0.78	0.71
	POS	0.923	0.892	0.907
	ALL	0.87	0.84	0.855

		Recall	**Precision**	**F-Measure**
Blog	**TOK**	0.98	0.966	0.973
	SF	0.642	0.764	0.697
	POS	0.92	0.872	0.895
	ALL	0.86	0.834	0.846

Tables 1(a-c): Performance Rates of the FST Module across the three genres where TOK is the tokenizer performance, SF is the tagger performance on identifying semantic features, POS is the POS tagging performance only and ALL is TOK+SF+POS

The FST module is robust in the sense that it is not affected by spelling variations as long as they are word-internal. The only relevant spelling variations for the FST module are those present in affixes and clitics such as the spelling variation in the future prefix of حـ */H/* vs. هـ */h/* or in the feminine marker clitic of ـة */p/* vs. ـه */h/*. Robustness is also reflected in the ability to tag borrowed words and EA-exclusive words as long as they abide by the same EA affix and clitic paradigm. Thus the sarcastic spelling of the borrowed word البورنامج */AlbwrnAmj/* (the show) is given the same tag as its more conventional spelling form البرنامج /*AlbrnAmj*/ (the show) – DT+SG_M_NN.

Borrowed words like بالابتوبات /*blAbtwbAt*/ (with the laptops) and تويتات /*twytAt*/ (tweets) are also successfully tagged as PRP+DT+PL_F_NN and PL_F_N, respectively. Similarly, EA-exclusive words شعوطنا */$EwTnA/* (he/it irritated us), هيدلعوها */hydlEwnA/* (they will pamper her), الجواز /*AlgwAz*/ (marriage) are successfully tagged as 3_SG_M_VBD+1_PL_OBJP, 3_PL_VBF+3_SG_F_OBJP and DT+SG_M_NN, respectively. More examples for the output of the FST module – taken from the most frequent 1000 words across the corpus – are given in tables (2) and (3).

Despite the robustness of the FST tagger, it does not work on problems like: labeling semantic features in the absence of morphological cues, contextual information to resolve ambiguities and morphologically-poor grammatical categories. The semantic features, especially of gender and number, are not always morphologically represented by affixes, which is the case in broken plurals, for example. Al-Sabbagh and Girju (2011) developed a corpus-based measure to resolve syntactic ambiguity using the degree of affiliation of a given word to each set of affixes and clitics being divided into noun-based, verb-based, adjective-based and adverb-based. However, this measure is not necessarily useful for highly ambiguous words such as nouns and adjectives, active participles and verbs such as عارف */EArf/* (I know). Finally, not all grammatical categories are equally morphologically productive in terms of affixes and clitics. Adverbs, for instance, are the least morphologically productive as they do not inflect for any of the semantic features of gender, number, tense or aspect and are only agglutinated to conjunctions. Thus they get the lowest performance results in our FST module: precision 0.61; recall 0.43 and F-measure 0.504.

Word	**POS Tag**
الجيش */Aljy$/* (the army)	DT+SG_M_NN
الثورة */Alvwrp/* (the revolution)	DT+SG_F_NN
خربتوا */xrbtwA/* (you destroyed)	2_PL_VBD
هيستحمرنا */htstHmrnA/* (he'll fool us)	3_SG_M_VBF+1_PL_OBJP
كويس */kwys/* (good)	SG_M_JJ
يالهوتي */yAlhwty/* (Oh my Goodness)	EXP
كمان */lmAn/* (too)	RB

Table 2: Sample FST Output Taken from the Most Frequent 1000 Words

For all the above reasons, the second phase of the hybrid-approach tagger involves manual annotations to fill in the gaps of the FST module and guarantee a gold-standard training corpus for the third phase, namely statistical modeling. Yet, the FST module saves manual annotations much work and enables tagging more training corpus and thus building more accurate statistical POS models for EA. The next two phases of the tagger and their performances are discussed in future work.

Our POS tagset consists of 45 tags – adapted from the Arabic Penn Treebank tagset – that combine into complex tag vectors, representing the semantic features, the morphological structure and the grammatical category of the target word. The sign (_) refers to semantic features of the grammatical category and (+) stands for the morphological boundaries between stems and their affixes and clitics. The number of unique combinations or unique tag vectors is 1,595 vectors.

Base-Phrase	Class	Description
الظابط الهربان *AlZAbT*/DT+SG_M_NN *AlhrbAn*/DT+SG_M_JJ (the fugitive officer)	ADJP	Adjectival Phrase
الفيديو بتاعك *Alfydyw*/DT+SG_M_NN *btAEk*/IN+PP$ (Your video)	NP	Free State *Idafa*: idafa is a syntactic structure in Arabic expressing a possession relation.
يا نهار اسود *yA*/VC *nhAr*/SG_M_NN *Aswd*/SG_M_JJ (What a horrible day!)	NP	Exclamation expression in a noun phrase syntactic structure
لما يحصل هجوم *lmA*/CN *yHSl*/3_SG_M_VBP *hjwm*/SG_M_N (When an attack happens)	VP	A verb phrase which is a part of a subordinate clause
اشتغالات المجلس العسكري *A$tgAlAt*/PL_F_NN *Almjls*/DT+SG_M_NN *AlEskry*/DT+SG_M_JJ (The military council tricks)	NP	*Idafa* structure
هو صعب أوي *hw*/SBJP *SEb*/SG_M_JJ >*wy*/RB (It/He's very hard,)	NP	Noun phrase that can map to a complete nominal (i.e. verbless) sentence in EA
يجنن أوي *yjnn*/3_SG_M_VBP >*wy*/RB (It is amazing.)	VP	Verb phrase

Table 3: Sample Base-Phrase Chunker Output

6. Base Phrase Chunker

Base-Phrase Chunking (BPC) is defined as a classification task with four classes: noun, verb, adjectival and adverbial phrases. Features used for classification include:

- **Semantic features** including gender, number definiteness agreement between subject and verbs, nouns and adjectives;
- **Morphological features** including subject and object clitics;
- **Syntactic feature including** transitive vs. intransitive verbs – verbs found to be encliticized to object pronouns are classified as transitive;
- **Lexical features**: using function words (i.e. prepositions, conjunctions, interjections, relative pronouns … etc.) as anchors;
- **Meta-linguistic features**: punctuation markers

Since BPC is built on the top of the FST tagger, its performance still needs improvements. Moreover, more features are to be included, especially statistical features. However, preliminary results are promising. Table (3) shows examples of the extracted base-phrases.

7. Corpus Information Extraction

After applying the threshold model of dialect identification, the total size of YADAC is 6M wordform tokens and 457K wordform types. It is distributed as 41% from online knowledge market services, 32% from microblogs and 27% from blogs and forums.

A Web interface to the corpus is to be made available in the coming few months at apfel.ai.uiuc.edu to enable users to extract such information as spelling variations of their search queries, morphological forms and their POS tags as well as base-phrases of which the original search query or any of its spelling/morphological variations are a part. Table (4) is an example of the returned search results for submitting ريس */rys/* as a query.

8. Related Work

In previous sections, we referred frequently to previous work on DA corpora and the current efforts to build them. In this section, we wrap up any points we have not covered in previous sections concerning related work.

McNeil and Faiza (2011) use traditional resources like books and newspaper articles and Web blogs and forums to build a 250k corpus of Tunisian Arabic to use is for building a bilingual Tunisian-English dictionary. According to McNeil and Faiza (2011), this is a non-trivial task given that Tunisian Arabic is mostly a spoken language since its native speakers prefer writing in French or in Arabic using Romanized script.

Zaidan and Callison-Burch (2011) use crowdsourcing to build a multi-dialectal Arabic corpus, in which native speakers judge the degree of dialectness of the readers' online commentaries on newspaper posts. Commentaries labeled as representing highly dialectal content are used to build a trigram based model to automatically identify

Original Search	Spelling Variation	Morphological Variations	Base-Phrases
ريس */rys/* (president)	رئيس */r}ys/* (president)	ريس */rys/* (president) → SG_M_NN رئيس */r}ys/* (president) → SG_M_NN الريس */Alrys/* (the president) → DT+SG_M_NN الرئيس */Alr}ys/* (the president) → DT+SG_M_NN للريس */llrys/* (to the president) → IN+DT+SG_M_NN رئيسهم */r}yshm/* (their president) → SG_M_N+3_PL_PP$ …	رجع ريس [VP]_*rjE*/3_SG_M_VBD *rys*/SG_M_N الريس بتاعنا [NP]_*Alrys*/DT+SG_M_N *btAEnA*/IN+1_PL_PP$ الرئيس المخلوع [ADJP]_*Alr}ys*/DT+SG_M_NN *AlmxlwE*/DT+SG_M_JJ …

Table 4: Sample Output for YADAC Information Extraction

Arabic dialects in the compiled comments. With an accuracy rate of 69.4%, their corpus size turns into 855k words of Modern Standard, Egyptian, Levantine and Gulf Arabic.

Diab et al. (2010) build COLABA, a multi-dialectal corpus based on blogs and forums, covering Egyptian, Levantine, Iraqi and Moroccan Arabic. The corpus offers linguistic analyses at many levels including morphological analysis, POS tagging and sentence boundary identification; all of which are semi-automatically performed. Moreover, the information retrieval engine of the corpus is to map Modern Standard Arabic queries to their dialectal equivalents. Currently, information about the performance rates for each of the linguistic analysis tasks is not available.

9. Conclusions and Future Work

This paper presented the first phase of YADAC – a multi-genre and multi-dialectal Arabic corpus. It incorporates data from multiple genres including microblogs, online knowledge market services and blogs. The first phase focuses on Egyptian Arabic, while later phases are to deal with Levantine and Gulf Arabic as spoken in the Arabian Peninsula. In addition to being among the largest current DA corpora, YADAC also offers linguistic analyses at the POS tagging and base phrase chunking.

There is much future work involved in improving YADAC. Extending work to other Arabic dialects is the first step. Improving performance of both POS tagging and base phrase chunking by reducing the input noise and incorporating more features are the second step. Finally, launching the information extraction Web tool to YADAC is another future step.

10. References

Alorifi, F. (2008). *Automatic Identification of Arabic Dialects Using Hidden Markov Models.* PhD Thesis, University of Pittsburg.

Al-Sabbagh, R. and Girju, R. (2011). Finite-State Transducer Tagger for Egyptian Arabic. Technical Report: Semantics Frontiers Group, UIUC.

Biadsy,F., Hirschberg, J. and Habash, N. (2009). Spoken Arabic Dialect Identification Using Phonotactic Modeling. *EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens

Buckwalter, T. (2002). *Arabic Morphological Analyzer (AraMorph).* Version 1.0.Linguistic Data Consortium, catalog No. LDC2002L49 and ISBN 1-58563-257-0

Canavan, A. and Zipperlen, G. (1996). CALLFRIEND Egyptian Arabic, LDC, Philadelphia

Canavan, A., Zipperlen, G. and Graff, D. (1997). *CALLHOME Egyptian Arabic Speech*. LDC, Philadelphia

Dasigi, P. and Diab, M. (2011). CODACT: Towards Identifying Orthographic Variants in Dialectal Arabic. *Proceedings of the 5th International Joint Conference on Natural Language Processing,* pages 318-326, Chaing Mai, Thailand, November 8-13, 2011

Diab, M., Habash, N., Rambow, O., Altantawy, M. and Benajiba, Y. (2010). COLABA: Arabic Dialect Annotation and Processing. *LREC Workshop on Semitic Language Processing,* pages 66-74, Malta, May 2010

Duh, K. and Kirchhoff, K. (2005). POS Tagging of Dialectal Arabic: A Minimally Supervised Approach. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages,* pages 55-62, Ann Arbor, June 2005

Habash, N., Diab, M. and Rambow, O. (2011). CODA: Conventional Orthography for Dialectal Arabic (CODA) Version 0.1 – July 2011. Technical Report: Center for Computational Learning Systems – Columbia University.

McNeil, K. and Faiza, M. (2011). Tunisian Arabic Corpus: Creating a Written Corpus of an Unwritten Language. *Workshop on Arabic Corpus Linguistics (WACL),* Lancaster University, 11-12 April 2011

Zaidan, O. and Callison-Burch, C. (2011). The Arabic Online Commentary Dataset: An Annotated Dataset of Informal Arabic with High Dialectal Content. *Proceedings of the 49th Annual Meeting of the ACL: Short Papers,* pages 37-41, Oregon, June 2011

Appendix (A): Sample of Gold-Standard Set for Dialect Identification

Corpus Item	Annotator #1	Annotator #2
محمود سعد رفض الظهور في حلقة اليوم من مصر النهاردة عشان شفيق هيظهر فيها ضيف	2	3
الرقم ده علشان لو أي حد عايز يبعت حاجه لأخواننا في ليبيا في عربيه طلعه بكرة	3	3
مش هوة لوحدة .. الوزير يوسف بطرس غالي اختفي هو كمان .. سافر برة مصر	3	2
يحصل كده مع واحدة عندها سنة يالهوي	3	3
لسه دلوقتي اتفضت المظاهرة،،،هى ابتدت من الضهر،،،بس ياريت بفايده ويلغوا المحاكمات العسكرية	3	3
اصلا واحد كان بيعاكس المعاكسة دى مرة وقفت كدا شوية م الصدمة بس مقدرتش اعمل حاجه	3	3
على الجزيرة واحد بيقول أنه محامي شهداء اسكندرية وتم منعه من الدخول وأن اللي جوة مش محامين الشهدا	2	3

Appendix (B): Sample of the POS tagset

POS tag	Meaning
DT	Determiner
SG	Singular
PL	Plural
F	Feminine
M	Masculine
2	2nd person
3	3rd person
VC	Vocative particle
CN	Conditional
IN	Preposition
SBJP	Subject Pronoun
OBJP	Object Pronoun
PP$	Possessive Pronoun
NN	Common Noun
JJ	Adjective
VBD	Past tense verb
VBP	Present tense verb
VBF	Future tense verb
RB	Adverb
EXP	Formulaic expression