ARLEX: A Large Scale Comprehensive Lexical Inventory for Modern Standard Arabic

Sawsan Alqahtani[1,3], Mona Diab[1], Wajdi Zaghouani[2]
[1]The George Washington University , [2]Hamad Bin Khalifa University, [3]Princess Nora Bint Abdulrahman University
[1]Washington DC, USA, [2]Doha, Qatar, [3]Riyadh, Saudi Arabia
{sawsanq, mtdiab}@gwu.edu, wzaghouani@hbku.edu.qa

Abstract

This paper introduces a lexical resource, ARLEX, for Modern Standard Arabic (MSA) that explicitly lists ambiguity at the lexical and syntactic levels for each token. Arabic orthography is known for being underspecified for short vowels and other markers such as letter doubling and glottal stops, known as diacritics. This leads to further ambiguity in orthography with real impact on natural language processing (NLP) applications, not to mention readability and human language processing. We specifically target listing alternative ambiguous forms of words within and across the same part of speech (POS), namely where tokens with no specified diacritics may have multiple possible diacritized alternatives. The entries in this dictionary are constrained to five POS tags: verbs, nouns, adjectives, adverbs, and prepositions. A morphological analyzer and disambiguator is leveraged to generate the desired linguistic properties. The resulting inventory, ARLEX, is a large scale comprehensive resource of words, recording their degree of ambiguity at various levels with example usages. ARLEX could be most useful for NLP applications, pedagogical applications, as well as socio- and psycho-linguistic studies.

Keywords: Lexicon, Diacritization, Arabic, Ambiguity

1. Introduction

Language ambiguity is an inherent characteristic of natural languages, which refers to the phenomenon where an instance can be interpreted in multiple ways. Ambiguity is at the core of problems faced by natural language processing (NLP) applications. Although humans have the ability to resolve such ambiguity based on their prior knowledge and context, there are instances (sentences, words, etc) that require multiple readings to resolve within context. The problem of natural language ambiguity is further exacerbated by conventional orthographic decisions where not all phonemes are explicitly represented.

Arabic standard orthography is one of these languages that is underspecified for such phonemes such as short vowels, gemination, etc, which are collectively represented as diacritic marks, aka diacritics. In other words, diacritics are crucial in denoting both pronunciations as well as meanings of such underspecified words. Most typical text in Arabic is rendered undiacritized, i.e. missing explicit diacritics, thereby compounding the linguistic ambiguity of text as observed, for instance, during the annotation of the various text types within the Qatar Arabic Language Bank project (Zaghouani et al., 2014; Zaghouani et al., 2015; Zaghouani et al., 2016b).

Orthographically fully specified Modern Standard Arabic (MSA) would consist of letters (consonants and long vowels) as well as diacritics. Diacritics can be divided into lexical, which specify the meanings of the words, and inflectional, which are added to provide syntactic roles of the words including syntactic case and mood endings as well as passivation. They comprise the short vowels (ُ ِ َ ْ),[1] gemination marks (ّ), nunation (ٍ ٌ ً) denoting indefiniteness markers, and the absence of vowels (ْ) typically used as a syllable delimiter as well as a mood marker. Because our concern is the meaning of the words, we only consider the internal diacritization (lexical) in this inventory and do not include syntactic case or mood diacritics nor general tanween (i.e. nunation) except where they are frozen, not syntactically motivated.[2]

A resource that lists words in their typical underspecified form and their corresponding possible meanings are useful for multiple purposes such as evaluating/building NLP tools, psycho-linguistic and socio-linguistic studies, as well as pedagogical applications.

In this paper, we present a monolingual large scale comprehensive lexical resource for MSA, ARLEX, which provides for each undiactrized word: various possible diacritized alternatives, together with other relevant information including: part of speech (POS), frequency of usage, genre usage, in addition to usage examples. It is a large scale automatically acquired inventory of words from multiple genres. The main objective of this inventory is to explicitly mark undiacritized forms of Arabic words when they are ambiguous. ARLEX represents different aspects of ambiguity at the word level: POS (syntactic level) and diacritized alternatives (lexical level). At the syntactic level, ambiguity indicates that the undiacritized word can have multiple possible POS tags. If an undiacritized word has a single POS then it is syntactically unambiguous. Within a given possible POS tag for an undiacritized form, a word may be lexically ambiguous as it may have multiple readings due to either multiple possible diacritizations or the same dia-

[1]We adopt the Buckwalter Transliteration (Buckwalter, 2002) system in the inventory.

[2]It is also worth noting that the diacritics may also include glottal stops, elongation, dots on letters, emphatic markers, or any additional normalization for the text such as replacing ا with آ, أ or إ where appropriate. However, we do not include them in this study.

critized form would have multiple meanings (similar to the bank 'financial institution' /bank 'river bank', in English). We account for all three ambiguity cases in ARLEX.

We use the morphological analyzer and disambiguator, MADAMIRA v1 (Pasha et al., 2014), to generate the desired features: POS, diacritized alternatives, and lemmas. It is important to note that ARLEX is not manually evaluated but rather uses human annotation in its development; crucially, it is tapping into the underlying morphological analyzer SAMA. Additionally, where available, we link entries in ARLEX with Tharwa (Diab et al., 2014). Tharwa enriches ARLEX diacritized lemmas with sense information as well as information such as meaning correspondents in dialects as well as English. Thus, ARLEX provides complementary information found in Tharwa and morphological analyzers such as SAMA.

2. Related Work

The phenomenon of language ambiguity has been investigated previously in several studies (Zaghouani et al., 2016; Versley, 2006). Zaghouani et al. (2016) provide linguistic analysis for possible ambiguity effects in MSA and show that automatic identification of ambiguous words helps reduce the annotation time. They ask annotators to tell whether they agree with the automatic ambiguity identification and then add missing diacritics to ambiguous words. Maamouri et al. (2012) created an educational tool and a corpus for Arabic reading enhancement by adding the diacritics to avoid the issue of word reading ambiguity.

In the the optimal diacritization scheme for Arabic orthographic representation (OptDiac) project (Bouamor et al., 2015; Zaghouani et al., 2016a), the focus was to create a large-scale annotated corpus with the diacritics for a variety of Arabic texts covering more than 10 genres to describe Arabic word pronunciation, and to create a valuable resource that can help address the issue of word reading ambiguity in the Arabic language.

Several lexical resources are available that help other research build and design their studies about languages (Zaghouani, 2014). This includes CELEX, Tharwa, AMPN, and SAMA. ARLEX is in line with such resources.

CELEX (Baayen et al., 1995) is a lexical resource that provides linguistic information for three languages: English (160,595 words), Dutch (381,292 words), and German (365,530). It compiles available manually annotated sources to provide detailed information about orthography, phonology, morphology, syntax, and frequencies at lemma and word levels. This resource is helpful for disambiguating the word forms since we may find multiple entries for the same word with slightly different information. ARLEX shares a subset of the objective presented in CELEX. CELEX, however, does not exist in Arabic.

Tharwa (Diab et al., 2014) is a multilingual lexicon that addresses the gap between different languages: English, MSA, and Arabic dialects with a current focus on Egyptian, Iraqi, Levantine. The publicly released Tharwa lexicon comprises 29,329 MSA, English, and Egyptian parallel instances. It is compiled to provide different linguistic information and help further studies in theoretical and computational linguistics. Although Tharwa provides a large repository of information about Arabic, it does not provide all possible alternatives for a given word as one of its objectives. The current proposed repository is an augmentation step to Tharwa where we link both resources using the index of MSA lemma and identify whether a diacritized lemma along with its POS has more than one sense.

AMNP (Hawwari et al., 2013; Zaghouani et al., 2016c) is a lexical semantic resource for Arabic morphological patterns. They built the morphological patterns' database using linguistic generalization of the semantic roles of the verbal predicates in the Arabic PropBank (Diab et al., 2008; Zaghouani et al., 2010; Zaghouani et al., 2012), which is a semantically annotated corpus of text from the Annahar Journal.

SAMA (Maamouri et al., 2010) is a morphological analyzer of MSA which provides all possible combinations of prefix, stem, and suffix for a given word. It also provides diacritization, clitic splitting information, and POS tags for each morpheme segment. SAMA maintains compatibility tables that show the appropriate combinations of prefix, stem, and suffix in MSA. This allows for the divination of all possible analyses for each given word. It includes 1,328 prefixes, 945 suffixes, and 79,318 stems. ARLEX is built on top of SAMA as MADAMIRA leverages it to provide all possible analyses and combinations as a first step in the disambiguation process. Our findings depend on SAMA output.

3. Dataset and Preprocessing

We use two datasets: the Arabic TreeBank (ATB) (Maamouri et al., 2008).[3] and the Contemporary Corpus of Arabic (CCA) (Al-Sulaiti and Atwell, 2006). ATB includes three genres: newswire (NW), broadcast news (BN), and web blogs (WB); CCA includes autobiography, children stories, economics, education, health medicine, interviews, politics, recipes, religion, science, short stories, sociology, spoken, and tourist travel. For preprocessing, we split all sentences in CCA at the punctation sentence periods.[4] Moreover, we leverage a dialectal identification tool, AIDA v2, to filter dialectal sentences (Al-Badrashiny et al., 2015), especially from the WB data in the ATB. Table 1 shows the number of sentences and words in the undiacritized forms for each genre, which also include numbers and punctuation. It is worth noting that ARLEX entries are in surface form as they occur in naturally occurring text with no preprocessing, which is different from SAMA and Tharwa where the entries are indexed by lemma form.

The ATB dataset provides human annotation for diacritics, POS tags, and lemmas for each undiaciritzed word. We use this information in our lexicon and complement them with automated information to construct a comprehensive lexicon as much as possible in terms of including all possible choices of alternative linguistic information. We apply the morphological analyzer and disambiguator, MADAMIRA (Pasha et al., 2014), to generate such alternatives for each

[3] Distributed by the Linguistic Data Consortium (LDC)

[4] CCA corpus includes long paragraphs; thus, we split the corpus by period which is the natural ending punctuation in most texts and fits our objective which is reducing the length of sentences.

Genre	# Sentences	# Words	Vocabulary Size
CCA	16,076	818,990	85,288
NW	23,488	630,634	65,404
BN	17,673	287,825	40,646
WB	3,818	58,468	18,222
TOTAL	61,055	1,795,917	142,381

Table 1: Corpus data statistics by genre indicating word types and word instances, where tokens are surface forms.

undiacritized word. MADAMIRA is trained on SAMA analyzer, discussed in Section 2., to retrieve all possible analyses for a given word and then uses a supervised classifier and a language model to rank the suggested choices. MADAMIRA do not provide analysis for words that are not recognized by its system; hence, we do not consider the automated analysis for such words. Table 2 shows some statistics of words with no provided analysis per genre.

Genre	# Types	# Words	% of No Analysis	% of Genre
CCA	7,311	41,334	60.17%	5.05%
NW	5,127	14,859	21.63%	2.36%
BN	1,702	10,356	15.07%	3.60%
WB	8,961	2,143	3.12%	3.67%
Overall	23,101	68,692	100%	3.82%

Table 2: This table shows the number of types (unique surface forms of words) and the number of words with no provided analysis in total and per genre. In addition, it shows the percentages of the words with no provided analysis compared to the total number of no-analysis words as well as the total number of words of the corresponding genre.

For CCA, we do not have human annotation for POS tags and lemmas, so we consider the top choices generated by MADAMIRA as the correct choice despite the possible errors (i.e. equivalent to the human annotation in ATB). CCA provides human annotated diacritization on the majority of the words which accounts for 93.64% of the data. However, where there is no human annotation for diacritization, we also use MADAMIRA's top choice.

For cleaning, we remove case and mood related diacritics from the diacritized version of the corpus since it does not contribute to the lexical meanings. We restrict our inventory to have a closed set of POS tags which are verbs, nouns, adverbs, adjectives, and prepositions.We do not accept any word normalization.[5]

MADAMIRA reports an accuracy of 95.9% for POS tagging and 86.3% in diacritization where both gold (i.e. humanly annotated) and automated words being compared have to be an exact match in tokenization, spelling, and full diacritization including syntactic case and mood markers. Alqahtani et al. (2016) evaluates MADAMIRA performance in diacritization in BN and WB genres, which are not used in MADAMIRA's training phase. They report 90.65% accuracy for full diacritization and 96.38% in full diacritization without case and mood diacritics which is the one we are using in the current work.

[5]MADAMIRA suggests alternative normalization variants for the following three groups (إ ء أ ا آ), (ى ي), (ة ه), as a result of anticipated spelling errors. For example, one of the suggestions for the word "أرجاء" [all around] in MADAMIRA is to convert it to "إرجاء" [postpone] which both have very different meanings.

4. Inventory

For each given undiacritized word in the corpus, we compile a list of relevant lexical information which are helpful in studies that concern lexical ambiguity in addition to potentially finding interesting relationships between ambiguity and other parameters. This lexicon is tab-serparated where each record contains the following information for each undiacritized word:

- `UNDIAC`: the surface word space-tokenized without any diacritic marks, i.e. undiacritized word (e.g. "الكتب" or "كتبه");
- `DIAC`: a possible diacritization for UNDIAC. (e.g. "الكُتُب" [the books] or "أكْتُبه" [I am writing it]);
- `UNDIAC and DIAC LENGTH`: the number of characters in UNDIAC and DIAC forms;
- `UNDIAC TOKEN`: the core token/stem of the word without any prefixes and suffixes (e.g. the stem "كتب" for both "الكتب" and "أكتبه");
- `DIAC TOKEN`: the diacritized version of the UNDIAC TOKEN. This is useful to group words that have the same underlying meanings (e.g. the stem "كُتُب" for "الكُتُب" and "أكْتُب" for "أكْتُبه");
- `LEMMA`: the diacritized lemma of the word. This is also helpful to further specify the meaning of the word (e.g. the lemma "كِتاب" for "الكُتُب" and the lemma "كَتَب" for "أكْتُبه");
- `POS`: the specific tag for DIAC as verb, noun, adjective, adverb, or preposition (e.g. "أكْتُبه" is a verb and "الكُتُب" is a noun);
- `AMBIG POS`: For each UNDIAC TOKEN, 0 means that there is only one possible POS tag, and 1 means that there are multiple possible tags (e.g. 1 for "كتب");
- `AMBIG DIAC WITHIN POS`: For each UNDIAC TOKEN within an associated POS, 1 means that we have multiple diacritic alternatives, and 0 means that there is only one possible reading (e.g. 1 for "كتب" as a verb);
- `AMBIG DIAC`: For each UNDIAC TOKEN, 1 means that we have multiple diacritized alternatives within and across POS, and 0 means that there is only one possible reading (e.g. 1 for "كتب");
- `Tharwa Index`: the index of Tharwa lexicon that has the same lemma and POS as the current instance;
- `Tharwa Ambiguity Within POS And Diac`: Tharwa lexicon includes possible senses of the diacritized lemma along with its POS represented as possible English translations. For each LEMMA instead of TOKEN, 0 means that there is only one sense for the word within the same POS and DIAC according to Tharwa lexicon, and 1 means that there are multiple possible senses (e.g. 0 for "كَتَب" and "كِتاب");

- `UNDIAC_*, DIAC_*, DIAC_POS_*:` The symbol * refers to the a specific genre. These labels include information about the frequencies of UNDIAC, DIAC, and DIAC within the associated POS in each genre, respectively. In calculating such frequencies, we only consider the top choices within context for each word as it occurs in the corresponding gold ATB sentences and the top ranked POS tags and diacritics generated by MADAMIRA for CCA. We do not take into consideration the other possible alternatives provided by MADAMIRA;
- `TOTALs:` this set of values corresponds to the total frequencies of UNDIAC, DIAC, and DIAC within an associated POS in the whole corpus;
- `SENTENCES:` Representative example sentences from the corresponding corpora, which show the associated DIAC and POS (top analysis) in context. It is important to note that some records do not have associated examples because MADAMIRA provides all possible alternative choices which may not be present in the corpus as a top choice. For ATB, we use the gold diacritics and POS tags as the top choice. For CCA, we use the gold diacritics where available, for cases missing diacritics, and for all POS tags, we use the top choice generated by MADAMIRA.

  Each record is unique in terms of the diacritic variant, lemma, POS, and diacritized token such that deeper linguistic layers are available to use for researchers. Because we are combining gold and automated resources, we need to obtain the linguistic information which is not provided by the human annotation from the corresponding automated analysis. This includes the same diacritics, POS, and lemma in case of ATB records and the same diacritics in case of CCA records. Thus, we compare the gold information with its automated counterparts; if there is a match, we accept the remaining linguistic information in the automated analysis. If there is no match, we try to maximize the mapping by editing the diacritized words in gold and automated resources so they match each other.

  For this reason, words that starts with 'وَال*' are considered the same as the ones start with 'وال*' so we do not consider the presence of the diacritic on the first letter, which is not necessarily specified since it can be inferred from the determiner 'ال'. We also filter out analysis that are exactly the same across all linguistic information except the diacritic in the prefixes 'ب' or 'ل' where the diacritic 'ِ' is optionally added; for instance, the set of words ('بِتَجَنُّب', 'بتَجَنُّب') and ('لِسُقُوط' , 'لسُقُوط') are the same across all linguistic information except the diacritic in the prefix. Furthermore, we filter those analyses whose undiacritized forms of the words are different than the corresponding gold undiacritized to ensure that there is no normalization of the word of any kind.

  Additionally, we filter gold entries of the following lemmas: typo (e.g. لِاسعارالنفط which is missing a space to separate two valid words), dialect (e.g. بِأيد), transerr (e.g. الجهاز), foreign (e.g. لِلنت), and DEFAULT (e.g. the invalid words *وَظفارن or *لِفوش) to further ensure the validity of the words and its associated examples. We are aware that some words of such types are valid but given incorrect lemmas[6] because of their presence with incorrect surrounding context so that annotators provide them a generic lemma. However, we choose to remove them all to make the process systematic and maintain valid lemmas to make the entries of this lexicon grouped in meaningful way.

  To link between Tharwa and ARLEX, We follow similar approach to maximize their mapping in terms of lemmas and POS tags. The POS tag sets differ in those resources in addition to the possible disagreement in the choice of POS tag for some words such as noun and adjectives. Thus, for matching, if the POS tag in ARLEX is noun, we consider identifying entries in Tharwa in a specific order of matching; we retrieve the entry of the first encountered POS tag and neglect the remaining choices. In particular, we compare the noun tag to the following order of POS tags in Tharwa: noun, vbn (i.e. verb, past participle), adjective, noun+prop (i.e. proper nouns), pple_act, noun_num (i.e. cardinal number), and noun+quant (i.e. quantifiers). If the POS tag is verb, we consider verb and then modal. If the POS tag is adjective then we consider adjective, pple_act, pple_pass, noun, and then adj+elative (i.e. adjective comparative). Adverbs and prepositions are mapped with their counterparts only because we have not encountered disagreement.

  We first identify Tharwa records in terms of lemma and POS information. If available, we add the associated Tharwa index to ARLEX. If no record is found, we compare a lemma and its POS tag of ARLEX to the word rather than the lemma and its POS in Tharwa . We similarly add the associated Tharwa index if available; otherwise, we consider the Tharwa index as '-1' which means no match.

4.1. Inventory Statistics

The analyses in our resource are augmented using possible valid combinations in SAMA. The number of total records in ARLEX is 343,919 where each instance is unique in terms of dicritized word, diacritized token, diacritized lemma, Tharwa index,and POS. The number of entries correspond to gold information is 155,495; the remaining are generated automatically without considering context.

4.1.1. Surface Forms of Words

As mentioned, we only consider a closed set of POS tags. Table 3 shows the percentages of considered and discarded tokens with respect to the POS tag per genre. As can be observed, the set of POS tags we have chosen constitute a large portion of the dataset which means that the discarded amount is not significant especially that it includes non-verbal words (e.g. numbers and punctuations).

The longest undiacritized surface word is 17 characters whereas the average length is 5 characters. Standard deviation for the undiacritized length is around 1.73 which shows that most words are spread over the average. We have 23 characters for the longest diacritized word and 8 characters on average. Most of the diacritized words in the corpus are of length near the average since the standard deviation is approximately 2.44.

Table 4 shows statistics at the surface forms in each category and overall corpus. The number of unique undiacritized surface words along with its POS is 148,396 whereas the number of unique undiacritized surface words

[6]For instance, the word 'الصَبْر' which has the lemma 'DEFAULT' in the ATB corpus and means 'patience or endurance'.

Statistic	NW	BN	WB	CCA	All
# Considered tokens	74.18%	81.42%	73.42%	69.81%	73.32%
# Discarded tokens	25.81%	18.58%	26.58%	30.18%	26.67%

Table 3: This table shows per genre: 1. The percentage of the surface forms of the words with the considered POS tags. 2: The percentage of the discarded surface forms of the words which have the remaining POS tags. Both combined construct the full dataset.

Statistic	verbs	nouns	adjs	advs	preps	Total
UNDIAC	37,098	79,451	30,903	171	773	113,570
Percentage	32.67%	69.96%	27.21%	0.15%	0.68%	-
DIAC	97,156	114,912	38,668	226	850	229,529
Percentage	42.33%	50.06%	16.85%	0.1%	0.37%	-
DIAC Increase	61.82%	30.86%	20.08%	24.34%	9.06%	50.52%

Table 4: This table shows the number of unique surface forms of the word for each category and for the whole corpus.

regardless of the POS is 113,570. This accounts for approximately 24% overlap between the surface words of different POS categories. Noun is the dominant category which accounts for 69.96% of the unique undiacritized words. Verb and adjective are the following categories which account for almost half of the occurrences of noun in the undiacritized version. Adverb and preposition comprise significantly a much smaller portion of the whole lexicon.

These observations change when the surface words are rendered diacritized. The number of unique diacritized surface word along with their corresponding POS is 481,341 whereas the number of unique diacritized words is 229,529 which shows around 53% overlap between diacritized words across different POS categories. Nouns and verbs are the most frequent POS tags that occur in the lexicon; adjective follow them in rank with a considerable gap. Verbs are the most frequent POS category that have diacritic variations which accounts for 61% increase of the number of surface words. The remaining POS categories experience increase due to diacritic variations at 20% to 30% except the preposition which goes under 9% increase. Overall, We have around 50% increase due to diacritic variations in the whole corpus.

4.1.2. Tokens and Lemmas

Table 5 shows statistics regarding the token and lemma levels for each POS tag. We consider the main token of the word to reduce sparseness in the data and to further focus in the underlying meaning. The number of undiacritized tokens is reduced by 44,330 which is 39% reduction compared to the surface forms. Noun is still the dominant POS category in the undiacritized token followed by verb and adjectives. The diacritized version of token follow the same pattern as the diacritized surface words. Verb is the most affected category due to diacritics which undergoes double increase in size. Adjective is the following POS category which accounts for around 33% diacritic variations. Noun then follows them in rank whereas adverb and preposition do not increase at a considerable percentage.

The lemma of the word further reduces sparseness as it focuses in the main meaning. As we can see from Table 5, noun has the most variations of lemmas followed by verbs and adjectives with significant gap. The number of unique lemma along with its POS tag is 28,606. The number of

Statistic	verbs	nouns	adjs	advs	preps	Total
UNDIAC	26,769	42,670	18,740	126	503	69,240
Percentage	38.66%	61.63%	27.07%	0.18%	0.73%	-
DIAC	58,733	60,568	28,030	143	537	134,652
Percentage	43.62%	44.98%	20.82%	0.11%	0.40%	-
DIAC Increase	54.42%	29.55%	33.14%	11.89%	6.33%	48.59%
LEMMA	6,736	15,653	5,725	113	379	25,223
Percentage	26.70%	62.06%	22.70%	0.45%	1.50%	-

Table 5: This table shows statistics at the token and lemma levels.

lemmas that are not found in Tharwa is 14,082 which account for 49.23%, which is a considerably high percentage. We also target ambiguity at different levels: syntactic and lexical. For the syntactic level, we have 18,043 undiacritized stems that are ambiguous at the POS which accounts for around 26%. For the diacritic alternatives, we have 25,664 undiacritized stems that are ambiguous in terms of the diacritic variations. The number of undiacritized tokens that are ambiguous within the POS tags are 26,067. If we do not constraint the ambiguity within the POS tag, the number of ambiguous words increase which is 39,413 undiacritized tokens. The number of lemmas along with their POS tags that are ambiguous within the diacritics is 1,395 such that we include all lemmas in our lexicon even the ones that have no link to Tharwa. We do not have knowledge of the sense ambiguity within the diacritics of the remaining lemmas.

4.2. Discussion

The absence of diacritics adds an additional layer of ambiguity in MSA. Diacritics help specify the exact meanings or even reduce the number of possible senses for a given undiacritized word. Although this sounds appealing and has proven beneficial in some tasks (Vergyri and Kirchhoff, 2004; AlHanai and Glass, 2014), full diacritization might also have performance degradation in some NLP applications (Alqahtani et al., 2016; Diab et al., 2007) and human reading speed.

Maamouri et al. (2006) shows that there are three types of ambiguity caused by diacritics: ambiguity within POS tags, ambiguity for the same grapheme without considering POS tags, and ambiguity that is related to case and mood information. The former type concerns structural and grammatical level of ambiguity whereas the first two types are lexical which is our focus in this paper.

It has been claimed that frequency may play a significant role in disambiguation where words that frequently occur tend to be less ambiguous and that such frequency varies depending on the genre (Stokoe et al., 2003; Mihalcea et al., 2004; Lee and Myaeng, 2002). The current resource provides three types of frequencies: diacritized within a particular POS, undiacritized, diacritized words in addition to fine-grained frequencies for each genre so that researchers would be able to pick certain genres suitable for their studies. This lexical resource shows gaps in the frequency distributions among the alternative choices for each undiacritized word which may lead to having multiple choices for the same undiacritized word that have equal or close frequency approximation. This leads to an erroneous expectations which one must be careful about when having a limited-size data.For example, the word 'أصغر' can have

the following valid choices: 'أَصْغَر'[lesser/minimum/less] or 'أَصَغِّر'[I make something smaller] with frequencies 11 and 0, respectively.[7] This example clearly shows the significant difference in the frequency.

POS is an important factor that specifies the syntactic category of a word in a sentence (Ballesteros and Croft, 1998). It further helps refine the available diacritic alternatives for the undiacritized word and identify the specific meaning. For example, the word 'شطر' can be 'شَطْر'[bisector] as a noun and ''[sundered] or 'شَطُرَ'[bisect] as verbs; there is no ambiguity within POS when the word is a noun because it can only take one form. On the other hand, the word is ambiguous in the case of verbs because it can take one of the two forms.

The main limitation of this resource is the automatic generation of linguistic information for each undiacritized word. In other words, we are relying on MADAMIRA for linguistic alternatives and have not evaluated this lexical resource through manual annotation. However, it is also costly and labor-intensive to create gold humanly-annotated lexical resource that provide all possible analysis and replace such a resource.

5. Conclusion

The main objective of this lexical resource is to help lexical-decision making based on explicitly marking within-POS ambiguity which means having multiple diacritic alternatives for the same undiacritized words within a particular POS. It also provides lexical information that is automatically generated including diacritic alternatives, POS, word length, frequencies (within and across varying corpora of different domains and genres) in addition to explicitly marking undiacritized words that have multiple possible POS, as well as providing usage examples. This resource will be used for readability experiments where we evaluate the impact of ambiguity and level of diacritization in human readings.

6. Bibliographical References

Al-Badrashiny, M., Elfardy, H., and Diab, M. (2015). Aida2: A hybrid approach for token and sentence level dialect identification in arabic. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 42–51.

Al-Sulaiti, L. and Atwell, E. S. (2006). The design of a corpus of contemporary arabic. *International Journal of Corpus Linguistics*, 11(2):135–171.

AlHanai, T. and Glass, J. (2014). Lexical modeling for arabic asr: A systematic approach. In *Proceedings of INTERSPEECH*.

Alqahtani, S., Ghoneim, M., and Diab, M. (2016). Investigating the impact of various partial diacritization schemes on arabic-english statistical machine translation. *AMTA 2016, Vol.*, page 191.

[7] In this example, we consider frequencies of the diacritized word constrained by a particular POS tag. 0 frequency means the surface form of the word never occurs in the corpus but it is a valid diacritic alternative for the word.

Ballesteros, L. and Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71. ACM.

Buckwalter, T. (2002). Arabic transliteration. *URL http://www.qamus.org/transliteration.htm*.

Diab, M., Ghoneim, M., and Habash, N. (2007). Arabic diacritization in the context of statistical machine translation. In *Proceedings of MT-Summit*.

Lee, Y.-B. and Myaeng, S. H. (2002). Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150. ACM.

Maamouri, M., Bies, A., and Kulick, S. (2006). Diacritization: A challenge to arabic treebank annotation and parsing.

Maamouri, M., Bies, A., and Kulick, S. (2008). Enhancing the arabic treebank: a collaborative effort toward new annotation guidelines. In *LREC*. Citeseer.

Maamouri, M., Zaghouani, W., Cavalli-Sforza, V., Graff, D., and Ciul, M. (2012). Developing aret: an nlp-based educational tool set for arabic reading enhancement. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 127–135. Association for Computational Linguistics.

Mihalcea, R., Tarau, P., and Figa, E. (2004). Pagerank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1126. Association for Computational Linguistics.

Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, volume 14, pages 1094–1101.

Stokoe, C., Oakes, M. P., and Tait, J. (2003). Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 159–166. ACM.

Vergyri, D. and Kirchhoff, K. (2004). Automatic diacritization of arabic for acoustic modeling in speech recognition. In *Proceedings of the workshop on computational approaches to Arabic script-based languages*, pages 66–73. Association for Computational Linguistics.

Versley, Y. (2006). Disagreement dissected: Vagueness as a source of ambiguity in nominal (co-) reference. In *Ambiguity in Anaphora Workshop Proceedings*, pages 83–89.

Zaghouani, W., Hawwari, A., Alqahtani, S., Bouamor, H., Ghoneim, M., Diab, M., and Oflazer, K. (2016). Using ambiguity detection to streamline linguistic annotation. *CL4LC 2016*, page 127.

Zaghouani, W. (2014). Critical survey of the freely available arabic corpora. *International Conference on Language Resources and Evaluation, OSACT Workshop*.

7. Language Resource References

Baayen, R., Piepenbrock, R., and Gulikers, L. (1995). Celex2 ldc96l14. In *Web Download. Linguistic Data Consortium.*

Bouamor, H., Zaghouani, W., Diab, M., Obeid, O., Oflazer, K., Ghoneim, M., and Hawwari, A. (2015). A pilot study on arabic multi-genre corpus diacritization. In *Arabic Natural Language Processing Workshop, Association for Computational Linguistics Conference*.

Diab, M., Mansouri, A., Palmer, M., Babko-Malaya, O., Zaghouani, W., Bies, A., and Maamouri, M. (2008). A pilot arabic propbank. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*.

Diab, M. T., Al-Badrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., and Eskander, R. (2014). Tharwa: A large scale dialectal arabic-standard arabic-english lexicon. In *LREC*, pages 3782–3789.

Hawwari, A., Zaghouani, W., O'Gorman, T., Badran, A., and Diab, M. (2013). Building a lexical semantic resource for arabic morphological patterns. In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, pages 1–6. IEEE.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., Bies, A., and Kulick, S. (2010). Standard arabic morphological analyzer (sama) version 3.1. *Linguistic Data Consortium, Catalog No.: LDC2010L01*.

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., and Palmer, M. (2010). The revised arabic propbank. In *Proceedings of the Association for Computational Linguistics Fourth Linguistic Annotation Workshop*, pages 222–226. Association for Computational Linguistics.

Zaghouani, W., Hawwari, A., and Diab, M. (2012). A pilot propbank annotation for quranic arabic. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature co-located with the North American Association Computational Linguistics conference (NAACL-HLT 2012*, page 78.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large scale Arabic error annotation: Guidelines and framework. In *International Conference on Language Resources and Evaluation (LREC 2014)*.

Zaghouani, W., Habash, N., Bouamor, H., Rozovskaya, A., Mohit, B., Heider, A., and Oflazer, K. (2015). Correction annotation for non-native Arabic texts: Guidelines and corpus. *Proceedings of The 9th Linguistic Annotation Workshop*, pages 129–139.

Zaghouani, W., Bouamor, H., Hawwari, A., Diab, M., Obeid, O., Ghoneim, M., Alqahtani, S., and Oflazer, K. (2016a). Guidelines and framework for a large scale arabic diacritized corpus. In *The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3637–3643. European Language Resources Association (ELRA).

Zaghouani, W., Habash, N., Obeid, O., Mohit, B., and Oflazer, K. (2016b). Building an Arabic Machine Translation Post-Edited Corpus: Guidelines and Annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Zaghouani, W., Hawwari, A., Diab, M., OGorman, T., and Badran, A. (2016c). Ampn: a semantic resource for arabic morphological patterns. *International Journal of Speech Technology*, 19(2):281–288.