Nabra: Syrian Arabic Dialects with Morphological Annotations

Amal Nayouf
Syrian Virtual University, Syria
amal_124724@svuonline.org

Tymaa Hasanain Hammouda
Birzeit University, Palestine
thammouda@birzeit.edu

Mustafa Jarrar
Birzeit University, Palestine
mjarrar@birzeit.edu

Fadi A. Zaraket, zfadi@utexas.edu
Doha Institute for Graduate Studies, Doha
American University of Beirut, Beirut

Mohamad-Bassam Kurdy
Syrian Virtual University, Syria
t_bkurdy@svuonline.org

Abstract

This paper presents Nâbīa (نَبْرَة), a corpora of Syrian Arabic dialects with morphological annotations. A team of Syrian natives collected more than $6K$ sentences containing about $60K$ words from several sources including social media posts, scripts of movies and series, lyrics of songs and local proverbs to build Nâbīa. Nâbīa covers several local Syrian dialects including those of Aleppo, Damascus, Deir-ezzur, Hama, Homs, Huran, Latakia, Mardin, Raqqah, and Suwayda. A team of nine annotators annotated the $60K$ tokens with full morphological annotations across sentence contexts. We trained the annotators to follow methodological annotation guidelines to ensure unique morpheme annotations, and normalized the annotations. F1 and κ agreement scores ranged between 74% and 98% across features, showing the excellent quality of Nâbīa annotations. Our corpora are open-source and publicly available as part of the Currasat portal https://sina.birzeit.edu/currasat.

1 Introduction

Dialectal Arabic (DA) content dominates informal writings in emails, social media, blogs, and social messaging. Interest in building computational resources for Arabic dialects has been in the rise to provide both (i) annotated corpora (Jarrar et al., 2022b; Alshargi et al., 2019; Khalifa et al., 2018; Bouamor et al., 2018; Jarrar et al., 2017; Al-Shargi et al., 2016; Zribi et al., 2015; Jarrar et al., 2014) and (ii) morphological dialect analyzers (Obeid et al., 2020; Khalifa et al., 2020; Pasha et al., 2014; Zribi et al., 2017; Abdul-Mageed et al., 2021).

In this paper, we present Nâbīa نَبْرَة, a set of corpora that complement existing Arabic dialect corpora by covering several dialect variants of Syrian Arabic. Nâbīa covers dialects from 10 Syrian localities including Aleppo, Damascus (a.k.a. Shami) , Deir-ezzur, Hama, Homs, Huran, Latakia, Mardin, Raqqah, and Suwayda. Nâbīa was collected from several sources including social media posts, scripts of movies and series, lyrics of songs, and local proverbs. Nine annotators worked on annotating 6K sentences with 60,021 tokens with full morphological annotations. Each word was annotated using: prefix(s), stem, and suffix(s), part of speech (POS), dialect lemma, MSA lemma, person, number, gender, gloss, and synonyms; in addition to the sub-dialect it belongs to.

ع/أداة مضارعة + ب/أداة مضارعة Particles for continuous present tense	**عبحكي** I am speaking
ع/أداة مضارعة + ب/أداة مضارعة + ت/للمضارع المخاطب المؤنث المفرد Feminine present second	**عبتتغاوي** You are being proud
مو/أداة نفي Negation particle	**مومشان** Not because

Figure 1: Examples of typical prefixes in Syrian dialects

ب/حرف جر + ك/ضمير متصل للمخاطب المؤنث المفرد Preposition + pronoun for singular feminine second person	**إشبك** What is wrong with you
ين/للمضارع: فاعله مخاطب مؤنث مفرد + ها/للمضارع: مفعوله غائب مؤنث مفرد Subject pronoun, present tense, feminine second person + object pronoun, present tense, feminine third person	**تحطينها** You put it

Figure 2: Examples of typical suffixes in Syrian dialects.

We adopted the same annotation methodology used to annotate the Palestinian Curras2 and the Lebanese Baladi corpora (Haff et al., 2022), which we also used with the four corpora of Lisan (Jarrar et al., 2023b). As we will discuss later, we adopted the SAMA tagsets (Maamouri et al., 2010), but we introduced new prefixes and suffixes that are commonly used in Syrian dialects (Figures 1 and 2).

1.1 Arabic and its Dialects

Over 300 million people speak Arabic, including Classical Arabic (CA), Modern Standard Arabic (MSA), and dialectal forms of Arabic (DA), in more than 23 countries. Natural language processing (NLP) research has traditionally focused on MSA because it is the most widely used form of Arabic in formal communication, newspapers, education, and media. CA dominates historical and cultural texts, whereas most colloquial and real-life communication uses local DA variants. DA content is lately gaining massive growth especially through blogs, social media, and local entertainment outlets in songs, movies, and series.

NLP pipelines often struggle with tasks involving DA content due to the inherent morphological richness of DA variants, their relative lack of resources compared to MSA, and the absence of a standardized orthography (Darwish et al., 2021). DA is classified regionally into Egyptian, Gulf, Levantine, North African, and Yemeni (Diab et al., 2010) with Syrian and Lebanese dialects considered as Northern Levantine, and Palestinian and Jordanian as Southern Levantine.

Syrian Arabic is well-understood across the Arab world due to its popularity in historical dramas, TV series, and soap operas. Twenty million Syrians speak it for daily life. Expatriates from the Levant (Jordan, Lebanon, Palestine, and Syria) helped spread the dialect throughout the world.

The rest of this paper is organized as follows. Section 2 reviews related work. We introduce Syrian as a Levantine dialect in Section 3 and discuss variant Syrian dialects in Section 4. Nâbra data collection and annotation methodology follow in Sections 5 and 6, respectively. We discuss the evaluation of Nâbra in Section 7, then we conclude in 8 and discuss limitations and ethics considerations.

2 Related work

There are several annotated corpora and lexicographic resources for MSA.

The LDC's Penn Arabic Treebank PATB (Maamouri et al., 2005) consists of about consists of 791,210 tokens collected from several news sources. PATB annotations include: tokenization, segmentation, POS tagging, lemmatization, diacritization, English gloss and syntactic structure. The LDC Ontonotes 5 (Weischedel et al., 2013) is another MSA corpus collected from news sources, consisting of about 330K tokens, which are annotated in the same way as the PATB. Ontonotes 5 also contains multiple layers of annotation, including the PATB annotation layer.

The Prague Arabic Dependency Treebank (Ar-PADT) (Hajič et al., 2004) is a treebank that contains morphological annotations for a corpus of MSA text. These annotations include lemmas, part-of-speech tags, and other morphological features. Ar-PADT contains about 224K words.

The LDC's SAMA is a stem database (Maamouri et al., 2010), which is an extension of BAMA (Buckwalter, 2004), designed only for morphological modeling. It contains stems and their lemmas and compatible affixes. It contains about 40K lemmas.

The lexicographic database at Birzeit University (Jarrar and Amayreh, 2019) provides a large set of MSA lemmas, word forms, and morphological features, which are linked with the Arabic Ontology (Jarrar, 2021) using the W3C LEMON model (Jarrar et al., 2019).

2.1 Dialectal Arabic Resources

There are several Arabic dialectal corpora with diverse morphological annotations.

An early pilot to build a Levantine Arabic Tree bank is presented in (Maamouri et al., 2006). The Palestinian dialect corpus Curras (Haff et al., 2022; Jarrar et al., 2017, 2014) comprises about $56K$ tokens. Each word in the Curras was annotated with different morphological features, including Prefixes, Stem, Suffixes, MSA lemma, Dialect Lemma, Gloss, POS, Gender, Number, and Aspect. The Lebanese Baladi corpus ($9.6K$ tokens) was developed in the same manner as Curras in order to form a more Levantine corpus (Haff et al., 2022).

CALLHOME (Canavan et al., 1997) is an Egyptian Arabic corpus with transcripts of telephone conversations in Egyptian. CALIMA (Maamouri et al., 2006) extended ECAL (Kilany et al., 2002) which built on CALLHOME to provide morphological analysis of Egyptian. The COLABA project (Diab et al., 2010) collected Egyptian and Levantine resources from online blogs leading to the construction of Egyptian Tree Bank (ARZATB) (Maamouri et al., 2014).

The Lisan (Jarrar et al., 2022b) consists of 1.2 million tokens, covering Iraqi, Yemeni, Sudanese, and Libyan dialects. The Yemeni corpus (about 1.05M tokens) was collected automatically from Twitter, while the other three dialects (about 50K tokens each) were manually collected from Facebook and YouTube. Each word in the four corpora was annotated with different morphological features, such as POS, stem, prefixes, suffixes, lemma, and a gloss in English.

A corpus of $200K$ tokens was morphologically annotated covering seven different Arabic dialects including Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi, and Moroccan (Alshargi et al., 2019). The GUMAR Emirati corpus (Khalifa et al., 2018) consists of 200K tokens collected from novels. MADAR (Bouamor et al., 2018) is an ongoing multi-dialect corpus covering 26 cities and their corresponding dialects. The Arabizi Tunisian corpus has 42K tokens (Gugliotta and Dinarelli, 2022).

The NADI (nuanced Arabic dialect identification) SharedTask (Abdul-Mageed et al., 2021, 2020) provided researchers with 10-million/21K unlabeled/labeled tweets and challenged researchers to identify the province-level dialects across 21 countries.

3 Syrian as a Levantine Dialect

The Levantine family of dialects can be linguistically split across the north including Lebanon and Syria, and the south including Palestine and Jordan. During the seventh century, Arabic spread across the area, which spoke Western Aramaic before then (Skaf, 2015).

Aramaic is a Semitic language continuum spoken during antiquity throughout the Levant where It served as the *lingua-franca*. Aramaic survives today through modern dialects such as Turoyo Syriac and Western Neo-Aramaic spoken in parts of Syria. It also survives more subtly in the noticeable substratum underlying Levantine dialects that differ from MSA on several linguistic characteristics such as phonology, syntax, morphology, and lexicon. This additionally motivates the development of morphologically annotated resources for Levantine dialects. In the sequel, we briefly review the differentiating factors between Levantine dialects, Syrian dialects, and MSA.

3.1 Levantine Phonology

Aramaic variants use the Abjad alphabet composed of 22 letters. When Arabic spread, the population of the region transcribed Arabic with its 28 letters using the 22-letter Abjad resulting in “Garshouni”, a Syriac writing tradition (Briquel Chatonnet, 2005). Adaptations to fit the additional letters led some Syriac graphemes to represent multiple phonemes of Arabic, especially some of the emphatic letters.

3.2 Syrian Phonology and Orthography

The Syrian Dialect has a glottal stop phoneme /ʔ/ that is cognate with either Hamza ء إ ا أ ؤ ئ /*ʾ* or Qaf ق /*q* . In spontaneous Syrian orthography, the two forms are distinguished in a manner similar to Lisan guidelines (Jarrar et al., 2023b). Exceptions include هلأ /*hlʾa* (now) written هلق /*hlq* in $Token$ with normalization rules to highlight its etymology link to هالوقت /*hālwqt* (this time). Less common spelling variations include devoicing ج /*ǧ* /ʒ/ to /ʃ/,which sometimes reflects in spontaneous orthography, e.g., نجتمع /*nǧtmʿ* /niʒtmiʔ/ (we meet) may appear as نشتمع /*nštmʿ* /niʃtimʕ/.

3.3 Levantine Morphology

Levantine inherits templatic morphology from Semitic languages where affixes play important roles. Several morphological differences exist when compared to MSA.

- Diacritic marking for syntax roles is less required in Levantine. They are marked with suffixes resulting in similar phonetic effects. For example, there is no need for writing Dhamma ـُ /*u* to distinguish the subject from the object. The MSA sentence غلب البطلُ الأسدَ /*ġlb ālbṭlu ālʾasda* (The hero conquered the lion) may switch the subject and object as in غلب الأسدَ البطلُ /*ġlb ālʾasda ālbṭlu* and the diacritics distinguish the roles. The Levantine variants are البطل غلب الأسد /*ālbṭl ġlb ālʾasd* and الأسد غلبو البطل /*ālʾasd ġlbw ālbṭl* (also written as الأسد غلبه البطل /*ālʾasd ġlbh ālbṭl*) with no need for diacritics.
- Some Levantine-specific morphemes do not exist in MSA such as عم /*ʿm* which denotes present continuous tense when it precedes imperfect verbs أنا عم باكل /*ʾanā ʿm bākl* (I am eating). Without it أنا باكل /*ʾanā bākl* means the general truth (I eat). MSA lacks such an indicator and the tense is inferred from context: أنا آكل /*ʾanā ʾākl* can mean both "I am eating” or "I eat”.
- Other morphemes include رح /*rḥ* and ح /*ḥ* that are Levantine future indicators compared to MSA’s س /*s* and سوف /*swf* . (iv) The progressive Levantine particle بـ /*b* (as in باكل /*bākl*) indicates imperfective verbs and no counterpart exists in MSA.

Syrian dialects lack the negation enclitic ش /*š* in a distinction from southern Levantine dialects. Syrian dialects make use of a number of future particles in free distribution. The progressive particle عم /*ʿm* strictly indicates active momentarily progression, while the progressive proclitic +ب /b indicates a wider habitual to the progressive range.

3.4 Levantine Dialect Lexicon

The Levantine lexicon is rich with loan words from other languages due to its cross-civilization frequent passage location.

Some Syrian words are originally Syriac, e.g., شوب /*šwb* (hot), or براني /*brāny* (outer). Other words are originally Turkish, e.g., دغري /*dġry* (straightforward). Some words encountered major semantic shifts, e.g., طز /*ṭz* comes from Turkish tuz for ‘salt’, then semantically shifted to mean ‘something unimportant’, and eventually ‘good riddance’. Other words were borrowed from French, e.g., ديكور /*dykwr* (decor) and جاتو /*ğātw* (gateaux), and from Persian, e.g., سرسري /*srsry* (badman). Military terms كورنيت /*kwrnyt* are used to specify accuracy and sharpness.

4 Variant Syrian Dialects

Syrian Arabic dialects are used in daily communication among most Syrians. Some of them are closer to Iraqi dialects, and the rest are closer to the Levantine southern Levantine dialects. Here, we review the most famous dialects spoken in Syria.

The Shami dialect is the dominant dialect in the Damascus area and is the most widespread and used Syrian dialect. As the dialect of the capital, it dominates Syrian series and films which are widely accepted, appreciated, and spread in the Arab world. It is used in dubbing and translation of foreign series (Turkish and Hindi).

Table 1 shows Shami dialect features:

- Sculpture: abbreviate two or more words.
- Substitution: an example is the replacement of ق /*q* with ء /*ʾ* hamza.
- Spatial inversion: the introduction or delay of letters to simplify pronunciation.
- Inclination: vowel exchange where ا /*ā* is pronounced ي /*y* .

The Aleppo dialect is dominant in Aleppo in northern Syria. It is distinctive in pronunciation and has a unique vocabulary used in Aleppo alone. The distinct vocabulary comes from ancient Syriac or Turkish. Examples of Syriac and Turkish vocabulary used in Aleppo follow. Syriac إيمت /*ʾiymt* replaces MSA متى /*mtā* (when), and Syriac دعك /*dʿk* replaces MSA عجن /*ʿğn* (knead). Turkish فرتيكة /*frtykh* and سكرتون /*skrtwn* replace MSA شوكة /*šwkh* (fork), خزانة /*ẖzānh* (closet), respectively.

Shami	MSA	Gloss	Rule
شو بدّك *šw bdk*	أي شيء بودّك *ʾay šyʾ bwdk*	what do you want?	النحت Sculpture
بالمشرمحي *bālmšrmḥy*	بكلام عربي واضح وفصيح *bklām ʿrby wāḍḥ wfṣyḥ*	In clear words	النحت Sculpture
أديش *ʾadyš*	كم يساوي *km ysāwy*	how much	ابدال Substitution
جوز *ğwz*	زَوْج *zawğ*	husband	قلب المكاني spatial inversion
هنيك *hnyk*	هناك *hnāk*	There	إمالة inclination

Table 1: Examples of Shami Dialect

With non-Arabic Syriac vowels (e, o), Aleppo words and verbs do not need the Dammah ـ (nourishing) and fatha ـ (accusative) diacritics. Verbs may require more than one object denoting the concept of تعدي /*tʿdy* (exceeds). Verbs connect to ن to denote the masculine plural instead of the MSA suffix م /*m* Turkish influence on Aleppo dialects morphs the pronunciation of fixed letters such as ج /*ğ* and ق /*q* to a majestic Turkish tone, and also reduces the pronunciation of vowels.

The Latakia dialect is spoken across the coast in Latakia and Tartous. It is a mixture of Arabic, Syriac, and Phoenician. It is characterized by the strong pronunciation of the letter ق /*q* , and also features the letter م /*m* before verbs to denote the present tense in all its forms, e.g.منكتب /*mnktb* (we write/are writing), ميدرس /*mydrs* (he studies/is studying).

The Raqqa dialect is one of the closest dialects to classical Arabic in terms of vocabulary. Raqqa enjoys a distinguished location on the shores of the Euphrates River. It is home (ديار /*dyār*) Mudar, who are Arabs from the north. Mudar were displaced to the Euphrates island several centuries before Islam. The Raqqa syllables sound commensurate to the corresponding classical Arabic syllables. For example, the pronunciation of کـ /*k* results in a thirsty ج /*ğ* as in کانت /*kānt* pronounced as جانت /*ğānt* . The letter ق /*q* is pronounced کـ /*k* similar to Yemeni dialects as in قاع /*qāʿ* (earth) pronounced as کاع /*kāʿ* .

The Deir-ezzur dialect aka. as الديرية /*āldyryh* is in proximity to the Euphrates as well, and preserves most of the phonetic aspects of standard Arabic. The significantly different phonemes are ق /*q* , ك /*k* and ء /*ʾ*, while there is no different in the gingival sounds.

The Homs dialect varies slightly across several rural and urban areas in the Homs district. This is mainly due to the habitual diversity of the countryside including a sizeable Turkman population.

This paper covers the dominant variant in the city of Homs. The Homs dialect is characterized by pronouncing the first letter in a word as if it has a Dammah ـُ /u diacritic (inclusion). This includes the name of the city حِمص /*ḥimṣ* , pronounced with a Kasra إِ /*i* dialect everywhere else. It also flips gender when it comes to masculine second-person إنتِ /*ʾinti* (you-male in Homsi) and feminine second person إنتَ /*ʾinta* (you-female in Homsi). It also differs in the pronunciation of the letter ج /*ǧ* as they phonetically annex a silent د /*d* resulting in a دج /*dǧ* sound.

The Hama dialect is spoken in the central Syrian governorates. It is a good representative of the Syrian Levantine dialects and close to the Shami one, as it tends to be soft and long in speech. It is distinguished by its eloquence and stretch in speech. Al-Hader (city in Hama) variant of the Hama dialect is the most prominent variant.

The Hauran dialect is spoken south of the Damascus countryside down to the Ajloun mountains in Jordan including Daraa. It is an ancient Arabic dialect spoken by multiple Arab tribes, where each of them has some distinguishing phonetic characteristics.

The Al-Suwayda dialect is spoken in Jabal al-Arab. The harshness of the mountain environment is reflected in the dialect's tone. It is taut, clear, and possesses a fast rhythm. Syllable notes exit soundly and eloquently. The concept of المضافة /*ālmḍāfh* played a major role in preserving the strength of the dialect. Therein, prominent, cultured, and experienced speakers exchange arguments. This highly contributed to the rigor of the dialect and brought it closer to standard and classical Arabic.

The Mardini dialect takes its name from the city of Mardin in الحسكة /*ālḥskh* . It is also called الجزراوية /*ālǧzrāwyh* in relevance to the الفراتية /*ālfrātyh* island. The dialect contains many Turkish, Persian, and Aramaic words.

5 Nâbīa Corpora Collection

We manually collected about 6,000 sentences with 60K tokens from Facebook, blogs, popular proverbs, Syrian films and series, local poetry, and lyrics of popular local songs in several Syrian dialects to build Nâbīa. Table 2 provides statistics on tokens, unique tokens, sentences, lemmas, nouns, verbs, and functional words in each of the 10 dialects Nâbīa covers.

The distribution relatively follows the order of dialect demographics. The Shami dialect is the richest with 17.3K tokens, used as primary dialect in Damascus, the capital, and in various Syrian TV series and films. Nâbīa contains 9.2K Aleppo tokens collected from popular stories on Facebook and from vocal poetry. Coastal Latakia features 7.9K tokens collected from film dialogues such as رسايل شفهية /*rsāyl šfhyh* - قمران وزيتونة /*qmrān wzytwnh* (Voice letters, Qumran and Zeitouna) and series such as ضيعة ضايعة /*ḍyʿh ḍāyʿh* (lost town). We also added common proverbs. Suwayda dialect features 3.2K tokens from the الخربة /*ālẖrbh* series. For Homs and Hama we collected jokes, and food discussions from social media blogs.

The Raqqa, Huran, and Mardin dialects feature the remaining 6.3K, 3.8K, and 1.6K tokens, respectively. We manually collected texts from social media for Raqqa and Huran. We found blogs documenting Raqqa. We used blogs and traditional stories for Raqqa, vocal poetry and lyrics of popular folklore songs for Mardini, and scenes from the Bedouin series for Huran dialects. We noticed that the collected data reflected spontaneous dialect documentation all across, contrary to what one would expect. Films and series were no less spontaneous than blogs and social media.

As Arabic is diacritic-sensitive (Jarrar et al., 2018), we did not remove any diacritics We tokenized the text of Nâbīa so that each token has a tuple with the following information.

⟨SentenceID, TokenID, TokenText, LocalDialectName, Governate⟩

6 Annotation Methodology and Features

We followed a semi-automated methodology, with an integrated productivity tool, friendly to non-programmers, to annotate Nâbīa.

6.1 Methodology

We developed the *Tawseem* annotation portal to help automate and validate the annotation process. The portal leverages spreadsheets, familiar to common users, and is powered by smart functionalities to improve annotation productivity. Figure 3 shows a snapshot of *Tawseem* annotation portal with the sentence شلون دا تدخلي تسلمي عالنفسا /*šlwn dā tdẖly tslmy ʿālnfsā* (how would you enter to greet someone in childbed).

For each token in the sentence, the portal saves 17 data elements. The $SentenceID$ and $TokenID$ columns identify the sentence and token.

Dialect لهجة	Damascus (Shami) شامية	Aleppo حلبية	Latakia ساحلية	Raqqa رقاوية	Deir-Ezzur ديرية	Homs حمصية	Huran حوران	Suwayda سويداء	Hama حموية	Mardin ماردلية
Tokens	17,274	9,255	7,893	6,284	4,322	4,139	3,807	3,150	2,322	1,575
Unique Tokens	7,123	4,452	3,829	3,389	2,453	2,047	2,094	1,681	1,355	949
Sentences	1,181	787	829	679	519	518	457	381	340	243
Unique MSA Lemma	4,230	2,825	2,548	2,367	1,909	1,543	1,580	1,312	1,051	686
Unique DA lemma	4,351	2,969	2,681	2,490	1,954	1,591	1,646	1,354	1,095	710
Nouns	7,700	4,251	3,771	3,316	2,384	2,064	2,090	1,527	1,135	694
Verbs	3,524	1,897	1,557	985	714	709	518	554	369	339
Functional Words	6,027	3,090	2,560	1,960	1,213	1,359	1,194	1,069	815	534

Table 2: Counts of tokens, unique tokens, sentences, unique MSA lemmas, unique dialectal lemmas, Nouns, Verbs, and functional words for each of the Syrian dialects

The rest of the columns specify the $rowToken$, $Token$, $prefix(s)$, $stem$, $suffix(s)$, POS, *gender*, *number*, *person*, *aspect*, *MSAlemma*, *dialectlemma*, $synonym(s)$, *gloss*, as well as the $sub - dialect$.

To simplify and accelerate the annotation process we leverage existing annotations in the following manner. First, we uploaded existing annotated corpora for dialects and MSA (Haff et al., 2022; Jarrar et al., 2023b) into the $Tawseem$ tools.

The tool allows the annotators to search and look up previous annotations. The lookup services search the database and return the top matching results ranked. Annotators can then select one of the results, and correct the corresponding features if needed.

Second, annotators can search the *Tawseem* portal annotations in other sentences whether made by themselves or by other annotators. This helps leverage previous annotations and improves the correction process. Additionally, annotators can look for existing annotations of a specific token in the *Tawseem* portal results.

6.2 Annotation Guidelines

Training annotators to use the *Tawseem* portal was straightforward as they were all familiar with the interface of a productivity spreadsheet. We also trained them with annotation guidelines for each of the features in Nâbra̅ as follows:

rowToken: $rawToken$ is the raw word as it appears in the corpus, without any modification.

Token : $Token$ is the normalized version of the $rawToken$. This entry corrects spelling errors if needed. The idea is to unify different forms of spelling the same word with one specification to mitigate the lack of spelling rules for Arabic dialects. It is necessary to unify the different ways one word can be written by multiple users to reflect the same pronunciation. We adopted the $Token$ guidelines used in the Lisan corpora (Jarrar et al., 2023b) as well as the Palestinian Curras2 and Lebanese Baladi corpora (Haff et al., 2022) so that Nâbra̅ can be included smoothly in a larger family of Arabic dialects for further research and applications if needed.

Dialect lemma (المدخلة المعجمية العامية) determines the dialect's original source of the token. Thus, if the word is a verb, we choose the past masculine 3rd person singular form as its colloquial origin. For nouns, we select the singular masculine, if not attained we select the singular feminine form. When introducing a new lemma, we specify the following: (i) definitions of senses in Arabic, which is important for word sense disambiguation tasks (Al-Hajj and Jarrar, 2021a; Jarrar et al., 2023a) and Word-in-Context WiC disambiguation tasks (Al-Hajj and Jarrar, 2021b). (ii) Equivalent lemmas in MSA (Jarrar et al., 2019, 2021).

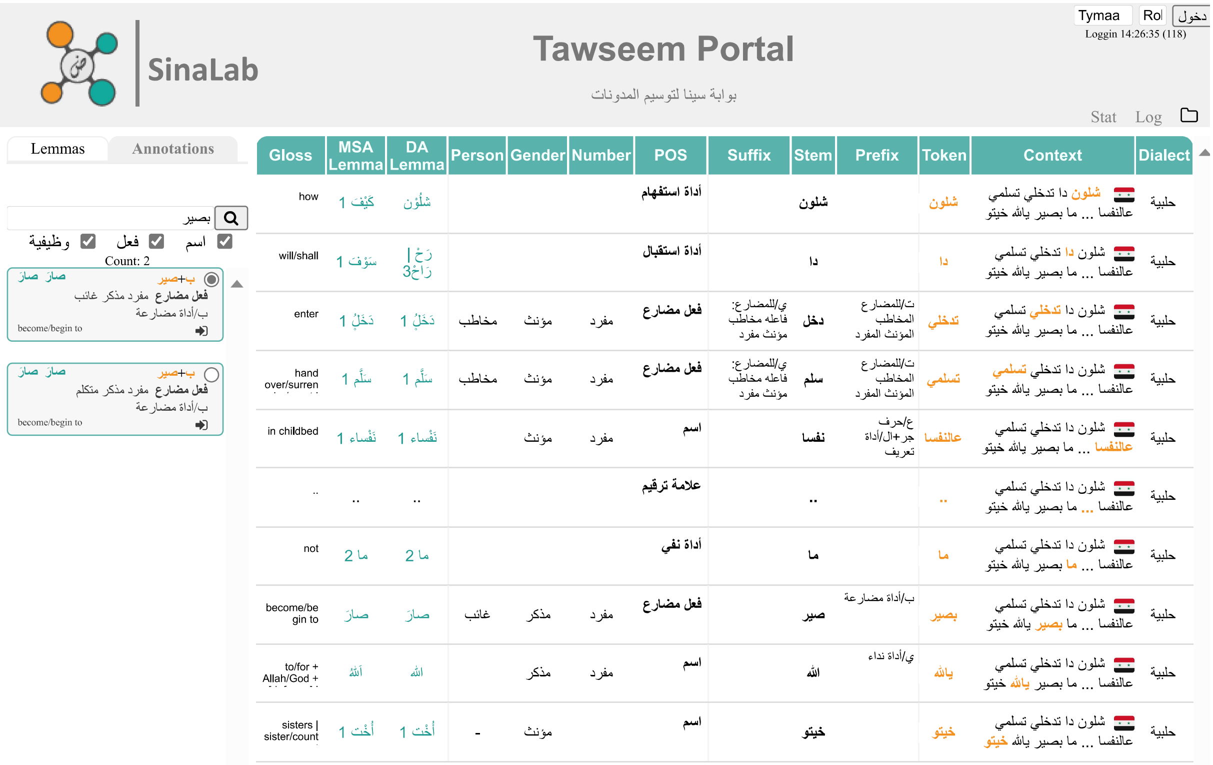
MSA Lemma (المدخلة المعجمية الفصحى) determines the MSA original source of the token. Table 3 shows examples of some tokens with their $Token$, and dialect and MSA lemmas.

The $Tawseem$ portal allows to search for lemmas in the Birzeit's Lexicographic database (Jarrar and Amayreh, 2019; Alhafi et al., 2019) and Arabic Ontology (Jarrar, 2021, 2011); otherwise, we introduced a new lemma.

The Synonym (المرادف) feature provides synonyms for the token and sometimes explains the token semantics. We used an online tool for automatic synonym discovery (Ghanem et al., 2023; Khallaf et al., 2023).

Gloss (المعنى بالانجليزية) specifies the meaning of the token in English. It typically specifies a short definition of lemma semantics. See an elaboration on the gloss formulation guidelines in (Jarrar, 2006).

POS (قسم الكلام) specifies the part of speech of the token. This concerns the grammatical category



بصير

☑ اسم ☑ فعل ☑ وظيفية

Count: 2

● ب+صير صارَ صارَ
فعل مضارع مفرد مذكر غائب
ب/أداة مضارعة
become/begin to

○ ب+صير صارَ صارَ
فعل مضارع مفرد مذكر متكلم
ب/أداة مضارعة
become/begin to

Gloss	MSA Lemma	DA Lemma	Person	Gender	Number	POS	Suffix	Stem	Prefix	Token	Context	Dialect
how	كَيْفَ 1	شلُوْن				أداة استفهام		شلون		شلون	🇸🇾 شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	حلبية
will/shall	سَوْفَ 1	رَحْ | رَاحْ3				أداة استقبال		دا		دا	🇸🇾 شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	حلبية
enter	دَخَلُ 1	دَخَلُ 1	مخاطب	مؤنث	مفرد	فعل مضارع	ي/للمضارع: فاعله مخاطب مؤنث مفرد	دخل	ت/للمضارع المخاطب المؤنث المفرد	تدخلي	🇸🇾 شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	حلبية
hand over/surren	سَلَّم 1	سَلَّم 1	مخاطب	مؤنث	مفرد	فعل مضارع	ي/للمضارع: فاعله مخاطب مؤنث مفرد	سلم	ت/للمضارع المخاطب المؤنث المفرد	تسلمي	🇸🇾 شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	حلبية
in childbed	نَفْساء 1	نَفْساء 1		مؤنث	مفرد	اسم		نفسا	ع/حرف جر+ال/أداة تعريف	عالنفسا	🇸🇾 شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	حلبية
..	..	..				علامة ترقيم		..		..	🇸🇾 شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	حلبية
not	ما 2	ما 2				أداة نفي		ما		ما	🇸🇾 شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	حلبية
become/begin to	صارَ	صارَ	غائب	مذكر	مفرد	فعل مضارع		صير	ب/أداة مضارعة	بصير	🇸🇾 شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	حلبية
to/for + Allah/God +	اَللّهُ	الله		مذكر	مفرد	اسم		الله	ي/أداة نداء	يالله	🇸🇾 شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	حلبية
sisters | sister/count	أُخْت 1	أُخْت 1	-	مؤنث		اسم		خيتو		خيتو	🇸🇾 شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	حلبية

Figure 3: Screenshot of the *Tawseem* annotation portal, our web-based annotation tool

of the token. We follow the SAMA tagset for compatibility reasons (Maamouri et al., 2010).

Stem (الجذر) specifies the segment of the token after removing suffixes and prefixes. It helps in the morphological analysis of the tokens. We follow the (Stem/POS) tagging schema used in (Maamouri et al., 2010) where the stem and POS are specified separated by '/'.

Affixes: prefixes and suffixes. We follow the prefixes السوابق and suffixes اللواحق tagging schema used in SAMA.

⟨Prefix1/POS⟩ + ⟨Prefix2/POS⟩ . . .

⟨Suffix1/POS⟩ + ⟨Suffix2/POS⟩ . . .

The schema specifies a sequence of affix and affix POS pairs separated by '+'. Each pair is an affix and affix POS separated by '/'.

Affixes and stems are morphemes where the concept of morpheme denotes the smallest morphological unit of text. Prefixes specify morphemes that connect to the beginning of a stem or to other prefixes to form a word. Suffixes specify morphemes that connect to the end of a stem or to other morphemes to form a word. Dialect affixes and their POS tags differ from MSA affixes and augment them due to the extended morpho-syntactic and semantic roles of dialect affixes.

Note here, for example, the synergy of using the future and progressive particles ع استقبال (FUT_PART) + ب مضارعة (PROG_PART) as prefixes to indicate present continuous tense for verbs in Aleppo as in عبشتغل /bštġl (I am working).

While most of the Syrian dialects precede present tense verbs with the IV1P POS with م مضارعة (PROG_PART), the Latakia coastal dialect applies it to almost all present tense verbs as with مأدرس /mʔadrs (I am studying). Latakia dialect also uses the prefix أ /ʔa for negation (and thus it corresponds to a NEG_PART POS tag) before present tense verbs as in أبعرف /ʔabʕrf (I don't know).

Person (الإسناد) specifies whether the subject of the token is a متكلم /mtklm (first), (مخاطب /mḫāṭb) (second) or غائب /ġāʔyb (absent) person when applicable.

Aspect (صيغة الفعل) concerns verbs and specifies whether they are in (مضارع /mḍārʕ) present for imperfective verbs (ماضي /māḍy) past for perfective verbs and (أمر /ʔamr) imperative tense.

Gender (الجنس) specifies whether a word is of مذكر /mḏkr male for masculine, مؤنث /mʔwnṯ female for feminine, or لا ينطبق /lā ynṭbq not applicable association when applicable.

Number (العدد) denotes مفرد /mfrd for singular, جمع /ǧmʕ for plural, مثنى /mṯnā for dual (to count two units), or لا ينطبق for uncountable words when

$rowToken$		$Token$	Dialect lemma	MSA lemma
ألت /ʾalt	I said	قلت /qlt	قال /qāl	قَالَ /qaāla
تختك /tẖtk	your bed	تختك /tẖtk	تخت /tẖt	سَرِيْر /saryr
مهندز /mhndz	engineer	مهندس /mhnds	مهندس /mhnds	مُهَنْدِس /muhandis
طريئ /ṭryʾy	street	طريق /ṭryq	طريق /ṭryq	طَرِيق /ṭariyq

Table 3: Example annotations for Nâbīa tokens

applicable.

7 Evaluation and Agreement

Before evaluating Nâbīa, we normalized the annotations to unify variant annotations that are equivalent. These variants occur due to human mistakes such as typos (ماصي /māṣy instead of ماضي /māḍy), ordering of tags in sequences of tags, and inconsistent use of separators and spacing.

Another source of variants is tokens with no feature values in the existing annotated dialects. Annotators have to come up with novel values. We detected these tag values, ranked them based on their frequencies, and clustered them based on their edit distance from each other. Then we reviewed them and unified them across Nâbīa and its features.

We developed a small suite of VBA scripts empowered with regular expressions to check for these variants and correct them automatically where possible. If automatic correction is not possible and human attention is required, then our reference annotators interfere to correct it.

7.1 Inter-annotation agreement

After the automatic corrections, six linguists visited the annotations to approve or correct them. This created a significant overlap of annotations as shown in Table 5. The overlap column shows the number of annotations per feature that had more than one annotation. Some of the second annotations were performed by the original annotator, so the reviewed column shows the number of annotations that were reviewed by two or more annotators. The unique column shows the number of unique values for the tokens with overlapping annotations.

The correction approach secured a significant overlap. We report the performance of the annotators in terms of precision, recall, and F1-score taking the correcting annotator as a reference in Table 4. A true positive (TP) for a feature value fv, denotes that the original annotation matched the reference annotation. A false positive (FP) for fv reflects an original annotator selecting fv for the token in conflict with the selection of the reference annotator. A false negative (FN) is when the original annotator fails to select fv for a token when the reference annotator selected it. Precision (P) and recall (R) are given by the ratios $TP/(TP + FP)$, and $TP/(TP + FN)$, respectively. The F1-score is given by $2PR/(P + R)$.

We also computed the Kappa-Cohen metric (McHugh, 2015) as implemented in the Scientific Kit Learn package (scikit learn, 2022). Table 4 shows the results where we compared the feature values of the reference annotators versus those of the original annotators.

The results show performance and agreement across all features. The κ scores are lower than the F-scores as the the κ metric accommodates for agreement by chance. The difference shows more with prefixes and suffixes as a significant part of the tokens had empty prefix and suffix, allowing more agreement by chance.

7.2 Qualitative Evaluation

To conduct a qualitative evaluation, we randomly selected about $7K$ annotations and reviewed them manually. We found a high agreement between the annotators who followed the specific guidelines and used our annotation tool. In what follows, we discuss some of the common mistakes:

(i) In rare cases, tokens specific to small local communities were hard to understand, Such as the token زنطر /znṭr (become cold) in the Latakia dialect. Although the annotators did their best to search external resources to understand such words, some mistakes still existed.

(ii) Tokens with no clear MSA equivalent led to difficulty in selecting MSA lemmas; thus, different annotators might not agree on selecting the same lemma. For example, the token عَمنوَّل /ʿamnwal may have several MSA lemmas, such as عام /ʿām (year), or ماضي /māḍy (past).

(iii) Semantic ambiguities in contexts led to disagreements on selecting lemmas. For instance, the token بقى /bqā has three possible meanings (was), (therefore) and (also). And sometimes all three fit the context.

8 Conclusion

This paper presents Nâbīa, a morphologically annotated corpora of Syrian Arabic dialects. The corpora contain about $60K$ tokens from 10 Syrian dialects, collected from social media platforms, movies and series, common proverbs, and song lyrics and poetry. To be compatible with SAMA

Feature	TP	FP	FN	P	R	F	κ
Stem	21,506	4,933	5,461	0.813	0.797	0.805	0.796
POS	20,727	2,979	3,316	0.874	0.862	0.868	0.843
Prefix	22,886	448	496	0.981	0.979	0.980	0.939
Suffix	22,096	1,247	1,380	0.947	0.941	0.944	0.837
DA Lemma	18,600	5,765	6,451	0.763	0.742	0.753	0.739
MSA Lemma	19,300	5,161	5,749	0.789	0.770	0.780	0.767

Table 4: Precision and recall results due to annotation correction with F and κ scores

Feature	Overlap	Reviewed	Unique
Stem	44,687	26,967	3,102
POS	39,007	24,043	56
Prefix	39,007	23382	163
Suffix	39,007	23,476	358
DALemma	41,579	25,052	3,586
MSALemma	41,579	25,050	3,352

Table 5: Reviewed overlap and unique feature values across Nâbr̄a

and other Arabic corpora, we chose to annotate the corpora using SAMA tagsets. To evaluate the quality of the corpora, we used the $F1$ and $kappa$ scores which show high agreement.

We plan to use Nâbr̄a to extend Wojood (Jarrar et al., 2022a; Liqreina et al., 2023) by annotating the corpora for Named Entity Recognition, similar to what we did with Curras and Baladi.

Limitations

The work in Nâbr̄a has the following limitations.

- Nâbr̄a covers 10 Syrian dialects. variants of these dialects and other smaller dialects confined in less urban localities exist. Future work should extend Nâbr̄a to better cover the Syrian dialect.
- Nâbr̄a addressed the Syrian dialects and their relation to the Arabic language and touched in prose on the relations to languages of origin such as Aramaic and Cyrillic. More data-oriented work is needed to relate Nâbr̄a to languages of origin that were spoken in Syria as well as to the geo-linguistic features of these languages.
- The annotation and evaluation process leveraged linguists who may be better at some of the dialects than others. We will make Nâbr̄a available online with correction suggestion capacities to accommodate for possible potential corrections.

Ethics Statement

The collection of texts used in Nâbr̄a respects intellectual property of the material. The annotation process employed annotators who were paid a fair rate per hour based on their living locality. Results from Nâbr̄a will be shared online for the research community to use and improve upon.

Acknowledgements

We would like to thank all the students who helped in the annotation process, especially Salam Eideh, Nagham Idrees, Shahd Khader, Shimaa Hamayel, Rwaa Assi, Hiba Zaid, Lama Saqqour, Ali Maihoob, and Eman Hariri.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Moustafa Al-Hajj and Mustafa Jarrar. 2021a. Arabglossbert: Fine-tuning bert on context-gloss pairs for wsd. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 40–48, Online. INCOMA Ltd.

Moustafa Al-Hajj and Mustafa Jarrar. 2021b. Lu-bzu at semeval-2021 task 2: Word2vec and lemma2vec performance in arabic word-in-context disambiguation. In *Proceedings of the 15th International Workshop* *on Semantic Evaluation (SemEval-2021)*, pages 748–755, Online. Association for Computational Linguistics.

Faisal Al-Shargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. Morphologically annotated corpora and morphological analyzers for Moroccan and sanaani yemeni Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1300–1306, Portorož, Slovenia. European Language Resources Association (ELRA).

Diana Alhafi, Anton Deik, and Mustafa Jarrar. 2019. Usability evaluation of lexicographic e-services. In *The 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEE.

Faisal Alshargi, Shahd Dibas, Sakhar Alkhereyf, Reem Faraj, Basmah Abdulkareem, Sane Yagi, Ouafaa Kacha, Nizar Habash, and Owen Rambow. 2019. Morphologically annotated corpora for seven Arabic dialects: Taizi, sanaani, najdi, jordanian, syrian, iraqi and Moroccan. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 137–147, Florence, Italy. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Françoise Briquel Chatonnet. 2005. De l'intérêt de l'étude du garshouni et des manuscrits écrits selon ce système. In *L'Orient Chrétien dans l'Empire musulman, en hommage au Professeur Gérard Troupeau*, Studia Arabica III, pages 463–475. Editions de Paris.

Tim Buckwalter. 2004. Buckwalter arabic morphological analyzer version 2.0. *LDC2004L02*.

Alexandra Canavan, George Zipperlen, and David Graff. 1997. Callhome egyptian arabic speech. *LDC97S45*.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab worlds. *Commun. ACM*, 64(4):72–81.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. Colaba: Arabic dialect annotation and processing. *LREC Workshop on Semitic Language Processing*, pages 66–74.

Sana Ghanem, Mustafa Jarrar, Radi Jarrar, and Ibrahim Bounhas. 2023. A benchmark and scoring algorithm for enriching arabic synonyms. In *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*, pages 215–222. Global Wordnet Association.

Elisa Gugliotta and Marco Dinarelli. 2022. TArC: Tunisian Arabish corpus, first complete release. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1125–1136, Marseille, France. European Language Resources Association.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a levantine corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Jan Hajič, Otakar Smrž, Zemanek Petr, Jaň Snaidauf, and Emanuel Beška. 2004. Prague arabic dependency treebank: development in data and tools. *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools.*

Mustafa Jarrar. 2006. Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In *Proceedings of the 15th international conference on World Wide Web (WWW2006)*, pages 497–503. ACM Press, New York, NY.

Mustafa Jarrar. 2011. Building a formal arabic ontology (invited paper). In *Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks*. ALECSO, Arab League.

Mustafa Jarrar. 2021. The arabic ontology - an arabic wordnet with ontologically clean content. *Applied Ontology Journal*, 16(1):1–26.

Mustafa Jarrar and Hamzeh Amayreh. 2019. An arabic-multilingual database with a lexicographic search engine. In *The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019)*, volume 11608 of *LNCS*, pages 234–246. Springer.

Mustafa Jarrar, Hamzeh Amayreh, and John P. McCrae. 2019. Representing arabic lexicons in lemon - a preliminary study. In *The 2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 2402, pages 29–33. CEUR Workshop Proceedings.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a corpus for palestinian arabic: a preliminary study. In *Proceedings of the EMNLP 2014, Workshop on Arabic Natural Language*, pages 18–27. Association For Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: An annotated corpus for the palestinian arabic dialect. *Journal Language Resources and Evaluation*, 51(3):745–775.

Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, and Khaled Shaalan. 2021. Extracting synonyms from bilingual dictionaries. In *Proceedings of the 11th International Global Wordnet Conference (GWC2021)*, pages 215–222. Global Wordnet Association.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022a. Wojood: Nested arabic named entity corpus and recognition using bert. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammad Khalilia. 2023a. Salma: Arabic sense-annotated corpus and wsd benchmarks. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*. ACL.

Mustafa Jarrar, Fadi Zaraket, Rami Asia, and Hamzeh Amayreh. 2018. Diacritic-based matching of arabic words. *ACM Asian and Low-Resource Language Information Processing*, 18(2):10:1–10:21.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023b. Lisan: Yemeni, irqi, libyan, and sudanese arabic dialect copora with morphological annotations. In *The 20th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE.

Mustafa Jarrar, Fadi A Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2022b. Lisan: Yemeni, iraqi, libyan, and sudanese arabic dialect copora with morphological annotations.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of emirati Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Salam Khalifa, Nasser Zalmout, and Nizar Habash. 2020. Morphological analysis and disambiguation for Gulf Arabic: The interplay between resources and methods. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3895–3904, Marseille, France. European Language Resources Association.

Nouran Khallaf, Elin Arfon, Mo El-Haj, Jon Morris, Dawn Knight, Paul Rayson, and Tymaa Hammoudaand Mustafa Jarrar. 2023. Open-source thesaurus development for under-resourced languages: a welsh case study.

Hanaa Kilany, H Gadalla, Howaida Arram, A Yacoub, Alaa El-Habashi, and C McLemore. 2002. Egyptian colloquial arabic lexicon. *LDC99L22*.

Haneen Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, and Muhammad Abdul-Mageed. 2023. Arabic fine-grained entity recognition. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*. ACL.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow, and Dalila Tabessi. 2006. Developing and using a pilot dialectal Arabic treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Hubert Jin, and Wigdan Mekki. 2005. Arabic treebank: Part 3 (full corpus) v 2.0. *LDC2005T20*.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic treebank: Impact of dialectal morphology on annotation and tool development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2348–2354, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mohamed Maamouri, David Graff, Basma Bouziri, Sondos Krouna, Ann Bies, and Seth Kulick. 2010. Ldc standard arabic morphological analyzer (sama) version 3.1. *LDC2010L01*.

Mary L. McHugh. 2015. Interrater reliability: the kappa statistic. *Biochemia medica*, 22.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland. European Language Resources Association (ELRA).

scikit learn. 2022. sklearn.metrics.cohen_kappa_score.

Roula Skaf. 2015. *Le morphème d= en araméen-syriaque : étude d'une polyfonctionalité à plusieurs échelles syntaxiques*. Theses, Université Sorbonne Paris Cité ; Università degli studi (Torino, Italia).

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0 ldc2013t19. *Philadelphia: Linguistic Data Consortium*.

Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith, and Philippe Blache. 2015. Spoken tunisian arabic corpus "stac": Transcription and annotation. *Res. Comput. Sci.*, 90:123–135.

Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith, and Philippe Blache. 2017. Morphological disambiguation of tunisian dialect. *Journal of King Saud University - Computer and Information Sciences*, 29(2):147–155. Arabic Natural Language Processing: Models, Systems and Applications.

A Appendix: Nâbr̄a Statistics

Table 6: Distribution of Gender feature. Arabic Words especially verbs and nouns and some of the functional words are annotated with “Male”“Female”. In some cases, the gender can be both, depending on the context, such as الجميع /*ālǧmyʿ* (everyone).

Gender	Count
Male	25,538
Female	11,790
Both	931

Table 7: Distribution of the Number feature. Arabic words especially verbs and nouns are annotated with “Singular”, “Dual”, “Plural”, and in some rare cases, the number can be "Any" like أبدى /*ʾabdā* (more important).

Number	Count
Singular	32,372
Dual	192
Plural	4,450
Any	163

Table 8: Distribution of the verbs’ Person: 1st person (متكلم), 2nd person (مخاطب), 3rd person (غائب).

Person	Count
1st	2,767
2nd	2,794
3rd	6,769

Table 9: Distribution of the POS tags and categories.

Category	POS	Count
NOUN Total: 28,932	NOUN	21,250
	ADJ	4,742
	NOUN_PROP	1,540
	NOUN_QUANT	556
	NOUN_NUM	315
	ADJ_COMP	257
	ADJ_NUM	152
	ABBREV	31
	DIGIT *	89
VERB Total: 11,166	IV	5,926
	PV	3,846
	CV	1,080
	IV_PASS	289
	PV_PASS	25
FUNC_WORD Total: 19,923	PUNC *	5,010
	PREP	3,133
	CONJ	2,506
	NEG_PART	1,642
	ADV	1,485
	PRON	1,252
	SUB_CONJ	991
	REL_PRON	687
	DEM_PRON	645
	INTERROG_PART	489
	VOC_PART	357
	PART	342
	PROG_PART *	218
	VERB	171
	INTERROG_PRON	166
	FUT_PART	130
	RESTRIC_PART	117
	FOREIGN	115
	PSEUDO_VERB	101
	EMOJI *	95
	VERB_PART	44
	INTERJ	43
	DET	40
	INTERROG_ADV	38
	EXCLAM_PRON	35
	FOCUS_PART	33
	PREP + SUB_CONJ	27
	REL_ADV	11
	Total	**60,021**