BRILL

JOURNAL OF JEWISH LANGUAGES 10 (2022) 24–53

brill.com/jjl

TAJA Corpus: Linguistically Tagged Written Algerian Judeo-Arabic Corpus

Ofra Tirosh-Becker | ORCID: 0000-0002-9234-1472
Professor of Hebrew and Arabic, The Hebrew University of Jerusalem, Jerusalem, Israel
otirosh@mail.huji.co.il

Oren M. Becker
Chairperson, Becker Consulting Ltd., Israel
becker.oren@gmail.com

Abstract

The Tagged Algerian Judeo-Arabic (TAJA) corpus is the first linguistically annotated corpus of any Judeo-Arabic dialect regardless of geography and period. The corpus is a genre-diverse collection of written Modern Algerian Judeo-Arabic texts, encompassing translations of the Bible and of liturgical texts, commentaries and original Judeo-Arabic books and journals. The TAJA corpus was manually annotated with parts-of-speech (POS) tags and detailed morphology tags. The goal of the new corpus is twofold. First, it preserves this endangered Judeo-Arabic language, expanding on previous fieldwork and going beyond the study of individual written texts. The corpus has already enabled us to make strides towards a grammar of written Algerian Judeo-Arabic. Second, this tagged corpus serves as a foundation for the development of Judeo-Arabic-specific Natural Language Processing (NLP) tools, which allow automatic POS tagging and morphological annotation of large collections of yet untapped texts in Algerian Judeo-Arabic and other Judeo-Arabic varieties.

Keywords

Judeo-Arabic – Algeria – corpus linguistics – linguistic tagging – digital humanities – natural language processing (NLP)

 | DOI:10.1163/22134638-BJA10020

1 Introduction[1]

For decades Arabic dialectology has been based on fieldwork in which the researcher interviews members of various local communities. The study of modern Judeo-Arabic (henceforth JA) dialects followed the same course, focusing on interviewing native speakers in their local townships and villages (e.g., M. Cohen 1912; Heath 2002). Over the course of the twentieth century this became increasingly difficult, especially in Algeria where the rising influence of French culture had eroded the status of JA, and the emigration of Algeria's Jews had dispersed this community. As a consequence of the diminishing Algerian JA speaking population, these dialects have become all but extinct, with only very old people who still remember them. Thus, customary linguistic field work is no longer a viable methodology for studying Algerian JA dialects. Fortunately, some Jewish communities in North Africa in general, and in Algeria in particular, were literarily very prolific and have left us vast collections of books and manuscripts written in JA. These are long lasting cultural treasures representing a variety of literary genres—Bible translations, translations of post-biblical texts, commentaries, liturgical texts, didactic texts, newspapers, and more. To date, only a select number of such written texts have been thoroughly analyzed from a linguistic perspective, while the vast majority have remained untouched. The new Tagged Algerian Judeo-Arabic (TAJA) corpus presented below preserves this endangered JA language and enables us to expand on previous dialectological work.

This article is organized as follows: After an introduction to North African and Algerian Judeo-Arabic (Section 1) we survey the many Arabic corpora and the very few Judeo-Arabic corpora that exist today (Section 2). Subsequently we present the TAJA corpus—its objectives, structure, tagging system, creation process, and statistical characteristics (Section 3). Examples of the utility of the TAJA corpus in linguistic studies are then detailed (Section 4). Finally, ongoing work towards developing Machine Learning morphology taggers for Algerian Judeo-Arabic based on TAJA is presented (Section 5) and directions for future research are discussed (Section 6).

1.1 *North African Judeo-Arabic*

Dialects spoken and written by the Jews of the Maghreb are collectively referred to as North African JA. Like many other Jewish languages, characteristics of North African JA include the use of Hebrew script, the presence of a Hebrew (and Aramaic) component, and co-existence of conservative traits,

1 This research was supported by an Israel Science Foundation grant 1191/18.

vernacular features, and heterogeneous elements. Following a few early works (Fleischer 1864; W. Marçais 1902), a cornerstone in the study of North African JA dialects was laid by M. Cohen (1912) in his important study of the JA dialect of Algiers. Since then, additional North African JA dialects have been studied and described. These have focused mainly on Moroccan JA dialects (e.g., Zafrani 1967; Heath 2002; Levy 2009; Chetrit 2016), and to a lesser extent on JA dialects from Tunisia (Bar-Asher 2005; Tedghi 2016), Libya (Yoda 2010), and Algeria (see below). Detailed descriptions of local dialects have been published for the JA dialects of Fes (Brunot & Malka 1939, 1940), Sefrou (Stillman 1988), and Tafilalet (Heath & Bar-Asher 1982) in Morocco, the JA dialects of Tunis (D. Cohen 1975), Sūsa (Saada 1956), and Gabes (Yoda 2006) in Tunisia, and the JA dialect of Tripoli (Yoda 2005), and Yefren (D'Anna 2021) in Libya. For an overview of North African JA see Tirosh-Becker (2021).

One aspect of North African JA that has been extensively studied is the substantial Hebrew (and Aramaic) component within these dialects, a characteristic common to many Jewish languages as well (Bunis 1993; Maman 2019). Many detailed discussions on the Hebrew component in Maghrebi JA in general, and North African JA dialects in particular, have been published (e.g., Bar-Asher 1992; Tedghi 2003; Henshke 2007; Chetrit 2010).[2] The presence of a Hebrew (and Aramaic) component in Judeo-Arabic is a synchronic manifestation of the diachronic process of language borrowing, which most likely originated from frequent code switching in Judeo-Arabic rabbinic discourse. Code switching, characteristic of bilingual speakers, indicates an alternation of two languages within a single discourse or even within a single sentence, which may be referred to as *insertional* code switching. With time, often following phonological and morphological adaptations, *insertional* code switching lexemes can become loanwords, which are an integral part of the absorbing language and are used by monolingual speakers as well (Poplack 1980: 583; Matras 2009:106–114).

1.2 *Algerian Judeo-Arabic*

Algerian JA is intriguing as it reflects a transition between the Moroccan dialectal area to its west and the Tunisian and Libyan dialectal area to its east. However, to date, it has not been studied as extensively as Moroccan JA. Detailed studies have been published for the JA dialects of Algiers (M. Cohen 1912), Constantine in eastern Algeria (Tirosh-Becker 1989, 2011a, 2014, 2019), and Gharadïa, a small oasis-dwelling community in the Sahara Desert (Tirosh-Becker 2015b, 2017; Bar-Asher 2017). Ahmed (2022) recently published an analysis of a sample

2 On code switching and borrowing in Moroccan Arabic see Heath 1989.

of JA letters written mostly by Algerian Jewish merchants during the late 18th century. Study of the Hebrew component in Algerian JA dialects has so far been limited to the dialects of Algiers (M. Cohen 1912) and of the western Algerian communities of Tlemcen and Aïn-Temouchent (Bar-Asher 1992). The study of North African JA dialects in general, and of Algerian JA in particular, benefits from the multiple dialectological studies of North African Muslim Arabic dialects, including detailed accounts of many Algerian Muslim dialects that were published by W. Marçais (1902, 1908), J. Cantineau (1936, 1937, 1938, 1940, 1941), P. Marçais (1936, 1947, 1954, 1956), and others (e.g., Mangion 1937; Ostoya-Delmas 1938; Grand'Henry 1972; Laraba 1981; Boucherit 2002). Also relevant to the study of Algerian JA are studies of the Muslim dialects of Tunisia (e.g., Stumme 1896; D. Cohen 1970; Talmoudi 1980; Singer 1984) and Morocco (e.g., Caubet 1993; Heath 2002), as well as broader accounts of North African Arabic dialects in general (e.g., P. Marçais 1977; Fischer & Jastrow 1980).

For centuries Algerian Jews lived in a state of multiglossia, using JA alongside Hebrew (the 'holy language'; *lešon ha-qodeš*) and being in contact with Muslim Arabic, Berber dialects, and other languages. However, following the 1830 French occupation of Algeria, the influence of French culture and language gradually increased. This process accelerated with the 1870 Crémieux Decree that granted French citizenship to most Algerian Jews, deepening their integration in the French experience and leading to the adoption of French as their main language of discourse, at the expense of the gradual weakening of local JA dialects (Tirosh-Becker 2015a:430–433). This process was most rapid in the capital, Algiers, and the prominent port city of Oran (Wahrān). Already in 1912, the dialectologist Marcel Cohen, who documented the JA dialect of Algiers, noted that JA was less prevalent than French among the younger Jews of Algiers (M. Cohen 1912). French influence was slower, yet present nonetheless, in landlocked conservative communities such as that of Constantine, Algeria's third largest city. The Jewish community of Constantine, which is situated in a mountainous region in eastern Algeria, remained a JA stronghold in the first half of the 20th century despite being the seat of one of the three French *consistoires* that managed Jewish life under French rule. Jewish presence in Algeria ceased with its independence in 1962, after which its Jewish population emigrated mainly to France, Israel, and Canada.

1.3 *Literary Genres in North African Judeo-Arabic*

JA was used not only for speech but also as a literary language. Some Jewish communities in North Africa were very prolific and have left vast collections of books and manuscripts written in JA, which are a long-lasting cultural treasure. These collections span an array of literary genres, ranging from JA poetry to a

variety of prose genres. Among the latter are JA translations of and commentaries on the Bible and some post-biblical texts (e.g., the Passover Haggadah and tractate ʾAvot), JA translations of modern prose, dictionaries, journals, and more. Hebrew was also used by Algerian Jews for a variety of literary genres, including Bible and post-biblical exegesis, halakhic literature, responsa, theological and ethical treatises, *derushim*, historical chronicles, and secular prose (Tirosh-Becker 2013).

An important literary genre, which has attracted a lot of attention, is that of the JA Bible translations known as *šurūḥ* (sg. *šarḥ*). As the medieval JA variety used by Rabbi Saʿadya Gaʾon in his monumental Bible translation (the *Tafsīr*) has become less intelligible throughout the centuries, newer JA Bible translation traditions emerged in Jewish communities throughout the Muslim world. *Šarḥ* traditions have also evolved for post-biblical Hebrew texts. Some of the *šarḥ* traditions from Morocco and Tunisia, which were orally transmitted, were later put down in writing and studied by modern scholars (Bar-Asher 1999, 2002; Maman 2000; Tedghi 2012).[3] From Algeria, we have *šurūḥ* from Constantine that were written down by Rabbi Yosef Renassia (e.g., Tirosh-Becker 1989, 2012). Rabbi Yosef Renassia (1879–1962), a prominent leader of the Jewish community in Constantine in the first half of the 20th century, published more than 100 volumes written in JA, encompassing Bible translations (*šurūḥ*) and commentaries, translations and commentaries of post-biblical texts, liturgical texts, translations of historiographic and halakhic books, dictionaries, grammar books, and more. This literary project—led and carried out by a single person—was the only one of its kind in 20th century Algeria. Some of the Constantinian JA translations of post-biblical Hebrew texts have been studied, among them JA translations of liturgical poems known as *Seliḥot* and *Hošaʿnot*, *Piyyuṭ Mi Khamokha*, and tractate *ʾAvot* of the Mishnah (Tirosh-Becker 2006, 2011b, 2011c, 2014).

Judeo-Arabic was also used in Algeria for journalistic writing. During the late 19th century (mainly 1885–1896) Jewish Algerian journals were either written in Judeo-Arabic or appeared as bilingual Judeo-Arabic and French publications, starting with the first Jewish journal published in North Africa, the 1870s bilingual French—Judeo-Arabic journal *ed-dziri* (Fr. *L'Israélite Algérien*). From 1896 through 1962, the year of Algeria's independence, which led to the emigration of Jews out of Algeria, almost all Algerian Jewish newspapers were composed

3 Among the Maghrebi *šurūḥ* to the Bible that have been studied to date are Rabbi Rephael Berdugo's *šarḥ Leshon Limmudim* (Morocco), Ḥai Diyyan's *Muqshiyya* (Tunisia), and the *šarḥ* of Issachar ben Susan al-Maghribi. For relevant bibliography see Maman 2000.

in French. A single exception is the journal *əl-Ḥikma*, which was published in Constantine in 1912–1913 and reappeared in 1922–1923 (Tirosh-Becker 2015a).

2 Judeo-Arabic and Arabic Corpora

In recent years, as methodologies of Digital Humanities (DH) became more widely available, digital corpora for numerous languages have been developed, with an aim to fuel progress in the field of Natural Language Processing (NLP). These digital corpora range from contemporary language (often extracted from the internet and digital media) to historical corpora, which rely on digitization of historical texts. Some are manually annotated or tagged with linguistic data, while most are unannotated and untagged. To the best of our knowledge, the TAJA corpus discussed herein is the only grammatically tagged digital corpus of *any* JA dialect.

2.1 *Judeo-Arabic Corpora*

The Judeo-Arabic Collection maintained by the Friedberg Jewish Manuscript Society (henceforth, the Friedberg JA Collection) is so far the only significant publicly available digitized JA corpus.[4] This large *unannotated* corpus holds approximately 4 million words from 110 *pre-modern* Judeo-Arabic texts from the 8th to the 18th centuries. The vast majority of the texts, however, are medieval, no later than the 13th century. These include seminal classical JA works, such as Rav Saʿadya Gaʾon's *Tafsīr* (Bible translation composed in 10th century Iraq), Maimonides' *Dalālāt al-Ḥāʾirīn* (*The Guide for the Perplexed*, a philosophical work composed in Egypt around 1190), and Judah Halevi's *Kitāb al Khazari* (*Sefer ha-Kuzari*, a medieval philosophical treatise composed in Andalusia around 1140).[5] This important corpus makes seminal JA works accessible for researchers and laymen alike. The corpus' web interface enables simple and composite searches of specific words or phrases across the corpus. As part of the digitization process, the words were manually annotated for language (either Arabic or Hebrew/Aramaic, most of which are quoted biblical verses), and this information is visually represented in the corpus web browser. This feature later enabled the development of an automated classifier for JA code switching, identifying code switching points between JA and Hebrew/Aramaic based on this corpus (Bar et al. 2015).

4 See https://fjms.genizah.org (last accessed May 5, 2022).

5 For the full list of texts that are included in the Friedberg JA corpus see https://vf.genizah.org/JA/FullBib/fullBib.htm (last accessed May 5, 2022).

The Friedberg JA corpus, however, is limited in several important aspects: (1) *No modern JA texts:* This corpus includes only pre-modern literary JA texts. The vast majority of the texts in this corpus are from the 8th–13th centuries, a period often denoted as the 'classical' era of JA culture. There is only minimal representation for later periods, and no text is later than the 18th century. In other words, although this corpus is a most valuable resource for *medieval* JA, it does not support the study of the multitude of *modern* Judeo-Arabic dialects. (2) *Limited genres*: The literary genres represented in this corpus are limited, consisting almost exclusively of scholarly literary texts. (3) *Limited geographical representation*: Most of the texts in the corpus are from Spain, Iraq, and to a lesser extent from the Levant and Egypt. There is a small representation for Yemen and only minute representation for North Africa.[6] (4) *No linguistic annotation*: Except for noting JA or Hebrew/Aramaic code switching, the texts of the Friedberg JA corpus are not annotated. In particular, there is no grammatical annotation of any kind in this corpus.

Ahmed (2018) reported on a subset of the Friedberg JA corpus which he re-annotated for JA / Hebrew code switching with a goal to investigate sociolinguistic aspects in medieval JA texts. This subset includes the first 100 pages of three medieval JA works (10th–12th centuries) chosen to represent a range of geographic, historical, and literary settings (a total of 300 pages, 67,000 words).[7] Plain digital versions of the texts were downloaded from the Friedberg JA Collection, without the limited code switching annotation already available in that corpus. The texts were then manually annotated using the Text Encoding Initiative (TEI P5) encoding structure, which is an encoding guideline established by the Text Encoding Initiative Consortium that uses the Extensible Markup Language (XML). The annotation distinguishes between inter-sentential code switching (i.e., code switching between sentences), intra-sentential code switching (i.e., code switching within the sentence boundary), borrowing, and Hebrew quotations.

Beyond the Friedberg JA Collection, the only additional digital JA corpus that we are aware of is the small publicly available corpus uploaded by Čéplö.[8] This corpus consists of a Libyan Šarḥ of Qohelet (Livorno 1897; 18,670 tokens),

6 The only 3 texts from North Africa included in the Friedberg corpus are the *Risāla* by Yehuda ʾIbn Quraysh (8th century; Algeria), vestiges of Rav Nissim Gaʾon's books (11th century; Tunisia), and Yosef ʾIbn ʿAqnīn's *Commentary on Song of Songs* (12th–13th centuries; Morocco).

7 The three works included are: Saadia ben Joseph al-Fayyumi's (10th century) *Kitāb al-Mukhtār fī l-ʾAmānāt wa-l-Iʿtiqādāt*; Moshe ben Jacob ibn Ezra's (11th–12th centuries) *Kitāb al-Muḥāḍara wal-Muḍākara*; and Yehuda Halevi's (11th–12th centuries) *al-Kitāb al-Kuzari*.

8 See https://www.bulbul.sk/crystal/#dashboard?corpname=judeo_arabic (last accessed May 5, 2022).

and a short *Mi Khamokha* poem in Maghrebi Judeo-Arabic (no further data on that poem; 894 tokens). The corpus is text only without annotation, tagging, or metadata of any sort.

2.2 *Arabic Language Corpora*

Large-scale digitized Arabic corpora, motivated by advances in the field of Natural Language Processing (NLP), are being developed at an increasing pace, although the vast majority of these corpora, such as the arTenTen corpus (Arts et al. 2014), are linguistically unannotated.[9] Most of these corpora focus on Modern Standard Arabic (MSA) and include texts that can be easily obtained from the internet and digital media (newswires, tweets, etc.), and only a small number of corpora focus on regional Arabic dialects (for reviews see, e.g., Zaghouani 2014; Shoufan & Al-Ameri 2015).[10] An up-to-date online catalogue of Arabic NLP datasets is maintained by the *Masader Project* and is available on Github.[11]

Among the relatively few linguistically annotated Arabic language corpora are the Penn Arabic Treebank (Maamouri et al. 2004) and the Prague Arabic Dependency Treebank (Hajič et al. 2004), both of which are linguistically annotated on multiple levels (part-of-speech, morphology, and syntax), following in the footsteps of earlier so-called "treebanks" (Nivre 2008). These corpora predominantly contain MSA news texts. The main annotated corpora of dialectal Arabic are the Egyptian Arabic treebank (Maamouri et al. 2014), which is annotated with morphological and syntactic information, and the similar but smaller Levantine Arabic treebank (Maamouri et al. 2006).[12]

As mentioned above, most digitized Arabic language corpora are linguistically unannotated. Primary examples of unannotated MSA corpora include ELRA's An-Nahar Newspaper Text Corpus and LDC's Arabic Gigaword corpus, whose 5th edition includes over 1 billion words in more than 3 million documents. The Open Source Arabic Corpus is a more diverse corpus covering genres such as sports, stories, and recipes (Saad & Ashour 2010). Several corpora include Classical Arabic (CA) texts, such as the King Saud University Corpus of Classical Arabic, which contains various texts from the first few

9 A chunk of this 5.8-billion-word corpus has been lemmatized and part-of-speech (POS) tagged with the MADA tool (Habash et al. 2009), but not manually tagged by linguists.

10 We thank Dr. Yonatan Belinkov for sharing his review of Arabic language corpora with us (private communication).

11 The Masader Project catalogue is accessible on https://arbml.github.io/masader/ (last accessed May 5, 2022).

12 Both Egyptian and Levantine treebanks are available by contacting the Linguistic Data Consortium (LDC).

centuries of Islam (Alrabiah et al. 2013); OpenITI which is a very large-scale diachronic corpus of Arabic (Romanov & Seydi 2019), and the cleaned and processed version of OpenITI prepared by Belinkov et al. (2019). Notable unannotated corpora of Arabic dialects include a corpus of Gulf, North African, Levantine, and Egyptian dialects (Almeman & Lee 2013); a corpus of North African, Egyptian, Levantine, Iraqi, and Gulf dialects based on user comments in online newspaper and Twitter tweets (Cotterell & Callison-Burch 2014); and the Arabic Online Commentary, covering Levantine, Gulf, and Egyptian dialects (Zaidan & Callison-Burch 2011).

In addition to the above listed Arabic corpora, there are several websites that provide online access only through a search interface (Al-Thubaity 2015; Alansary et al. 2007, and the Leeds Arabic Internet Corpus[13]). Web-based dialectal Arabic corpora include the Tunisian Arabic Corpus that contains texts from diverse sources such as folktales, screenplays, web forums, and transcribed recordings; and the Gumar corpus of Gulf Arabic that has semi-automatic morphological annotation (Khalifa et al. 2016).

2.3 *Corpora of Algerian Arabic*

Very few corpora of Algerian dialects have been reported so far. The only annotated Algerian Arabic dataset is the small NArabizi (Algerian Arabic written in Latin script) treebank presented in Seddah et al. (2020), which comprises approximately 1,500 sentences (22,000 words). Most of them (1,300 sentences) were extracted from an Algerian newspaper's web forum, and the rest are lyrics of songs collected manually from the web. These sentences were manually annotated for part-of-speech (POS), morphology (gender, number, tense, and verbal mood), code switching, and syntax, and are accompanied by a translation into French.

Among the few unannotated Algeria corpora is PADIC, a Parallel Arabic Dialect Corpus, which includes 6,400 sentences each in multiple language varieties: MSA and in five Arabic dialects, two of which are Algerian (Annaba, Algiers) and one is Tunisian (Sfax) (Meftouh et al. 2015, Harrat et al. 2015). Another is CALYOU, Comparable spoken ALgerian extracted from YOUtube (5,190 sentences), which attempts to match Algerian themed YouTube comments written in an Algerian dialect, MSA, and French (Abidi et al. 2017). Finally, KalamDZ, a corpus of recorded speech in a variety of Algerian dialects, encompassing 104 hours of recorded speech from YouTube and online radio, was recently published by Bougrine et al (2017).

13 Accessible at http://corpus.leeds.ac.uk/internet.html (last accessed May 5, 2022).

3 The TAJA Corpus

3.1 *Corpus Characteristics*

The Tagged Algerian Judeo-Arabic (TAJA) corpus was created as a *linguistically annotated* digital corpus of *genre-diverse* texts in *modern written* Algerian JA. Its goal is both to preserve this endangered JA language that may fall into oblivion, as well as augment and expand on previous dialectological work that was based on *oral* fieldwork or analysis of a limited number of carefully selected written texts. Hence, the main features of the TAJA corpus are as follows:

- *Modern JA*—The texts included in this corpus were printed in the late 19th and early 20th centuries. It should be noted, however, that some of these texts reflect older orally transmitted translation traditions that, although written down in the 20th century, reflect language traditions that may date back a few centuries.
- *Written texts*—Unlike fieldwork with informants, which was characteristic of JA dialectology, this corpus is based on written JA texts published in Algeria itself, or published by Algerian authors in foreign printing houses, such as those in neighboring Tunisia.
- *Genre-diversity*—In creating TAJA we directed special attention to ensure genre-diversity within the corpus. A rich representation of diverse literary genres and language registers is more likely to capture, within the finite limits of the corpus, the richness and diversity of the language itself. Indeed, a key limitation of corpora is often the restricted scope of the incorporated literary genres, whether they are based on digital media such as the Penn Arabic Treebank (newswires) or text-based such as the Friedberg JA corpus discussed above. In TAJA we aimed to balance traditional JA Bible translations (known as *šurūḥ*), whose language is known to be conservative, with JA newspapers and contemporary writings that use a more vernacular variety. We also aimed to balance JA translations of Hebrew texts, which are influenced by the source language, with original free writings composed in JA.
- *Linguistic annotation*—The goal of the TAJA corpus is first and foremost to support linguistic research of Algerian JA, with an eye on future interaction with Natural Language Processing (NLP) and the development of JA-specific NLP tools. Multi-level linguistic annotation of the corpus is of course a prerequisite to meet these goals. Hence, significant effort was dedicated to manually annotating the TAJA corpus with rich, high-quality linguistic tags across multiple levels, including part-of-speech (e.g., *noun, verb, adjective,* etc.) and morphology (including *lemma, verbal stems, tense, person, number, gender*, as appropriate for different grammatical categories, and *affixes*). The annotation was done by expert JA linguists.

3.2 *Corpus Selection*

The TAJA corpus includes a collection of modern JA texts published in Algeria in the late 19th and early 20th centuries. These texts were selected to represent an array of prose genres written in JA by Algerian Jews. These texts were classified into five genre-groups:

(i) JA translations of the Hebrew Bible, known as *šurūḥ*.
(ii) JA translations of seminal post-biblical Hebrew texts (such as the Mishnah, the Passover Haggadah, and liturgical poems known as *piyyuṭim*), which are often also referred to as *šurūḥ*.
(iii) Other JA translations.
(iv) Original writings composed in Algerian JA, such as commentaries, sermons, and religious texts.
(v) Newspapers written in Algerian JA.

A summary of the texts and genres that comprise the TAJA corpus is given in Table 1.[14] Currently, the TAJA corpus includes 17 texts, and more than 63,000

TABLE 1 The linguistically annotated Tagged Algerian Judeo-Arabic (TAJA) corpus

#	Source	Genre
1	JA translation of Psalms	JA Bible translation (*šarḥ*)
2	JA translation of Proverbs	JA Bible translation (*šarḥ*)
3	JA translation of Joshua	JA Bible translation (*šarḥ*)
4	JA translation of the Hafṭarot	JA Bible translation (*šarḥ*)
5	JA translation of Passover Haggadah	JA Translation, post-biblical
6	JA translation of the Hosha'anot	JA Translation, post-biblical
7	JA translation of the Seliḥot	JA Translation, post-biblical
8	JA translation of *Mishnah 'Avot*	JA Translation, post-biblical
9	JA translation of Maimonides' *Mishne-Torah, Hilkhot sekhirut*	Other JA translations
10	JA commentary on the Hosha'anot	Original writings in JA
11	JA commentary on Joshua	Original writings in JA
12	JA commentary on *Mishnah 'Avot*	Original writings in JA
13	JA commentary on the Seliḥot	Original writings in JA
14	The book *Peraḥ Shoshan* by Shalom Bekache	Original writings in JA
15	The book *'Or 'Olam* by Shalom Bekache	Original writings in JA
16	JA journal *əl-Ḥikma*	JA journalism
17	JA journal *Maguid Micharim*	JA journalism

14 Full references for the specific texts are detailed at the end of this paper.

annotated Algerian JA words. TAJA continues to grow as additional annotated texts are gradually added to it.

3.3 *Corpus Structure: Word-based Digital Records*

The basic structural elements of this digital corpus are the individual words, stored in word-based digital records. Namely, each word in each text is associated with a unique digital record, denoted herein as *word-record*. The role of the *word-record* is to place the word in the tree-like context that flows from the full text (which can still be accessed), down through the sentence-level, to the word-level, and finally to the morpheme-level. The *word-record* connects each specific word 'upwards' to its context, i.e., the sentence and the full text, as well as 'downwards' to its grammatical components via multiple layers of annotation (see examples in Table 2). The words and context are stored in their original Hebrew script.

Each word-record includes the following elements:

1. *Word*—The word exactly as it appears in the text, including pronominal suffixes, possessive pronouns, definite article, conjunctives, etc. For example, וקאללהם (*u-qāl-l-hum*, 'and [he] told them'), פאליל (*f-əl-lēl*, 'in the night').
2. *Word-core*—The basic nominal form of the word without affixes, pronouns, etc. or the root of a verb, e.g., וקאללהם → קול (√*qwl,* 'to tell'), פאליל → ליל (*lēl*, 'night').
3. *Context*—The complete sentence in which the word appears, enabling analysis of syntax and morpho-syntax. It is repeated for every word in that sentence.
4. *Text metadata*—A shorthand pointer to the full reference of the text in which the word occurs, e.g., the pointer *Gn_Psalms_44:5* stands for the JA translation (*šarḥ*) of Psalms 44:5 published by Rabbi Yosef Renassia (Hb. גנאסיה; full references are listed at the end of this article).
5. *POS annotation*—part-of-speech tagging, for details see below.
6. *Morphology annotation*—morphology tagging, for details see below.[15]
7. *Linguistic comments*—free-text notes that address a broad range of phenomena, including assimilation, dissimilation, spread of emphatic pronunciation, metathesis, comments on script, secondary roots, Hebrew influence, internal passive voice, special syntax issues, word order, semantic shifts, colloquial vs. archaic features, alternative translations, and more.

15 Within the morphology tags we also tagged loanwords, either as Hebrew (and Aramaic) or as Romance (French, Spanish, Italian).

TABLE 2 The structure of a *Word-record* with a specific example

Word*	Word-core*	
צולטאני (*ṣulṭāṇ-i*)	צולטאן (*ṣulṭāṇ*)	
Text metadata	**Context***	
Gn_Psalms_44:5 (= JA translation of Psalms 44:5 published by Rabbi Yosef Renassia)	אנתא הווא צולטאני יא אללאה וצי מג'יתאת יעקב: *enta huwa ṣulṭāṇ-i ya aḷḷah* *waṣṣi mġītāt yaʿakov* (= You are my King, O Lord, command the salvations of Jacob.)	
Part-of-speech	**Morphology tags**	**Linguistic comments**
Noun	Sg + Pronominal Suffix 1Sg	1. *waw* indicates vowel quality. 2. The Hebrew letter צ reflects emphatic spreading.

*The words and context are stored in the *word-record* only in their original Hebrew script; the transcription and translation are added in this example for clarity.

The digital corpus and its word-records are currently implemented as an Excel database using Perl programs and embedded macros. We are planning to make it broadly available online at a future date.

3.4 *Corpus Generation Process*

The corpus generation process required converting all the selected texts to digital form and then having expert JA linguists linguistically annotate them. This process spanned several years and was carried out with the diligent help of more than a dozen research assistants (RAs). Undergraduate level RAs were tasked with digitization, and graduate level RAs (pursuing advanced studies in JA) carried out the linguistic annotation. Both digitization and annotation

included proofreading. The entire process was supervised by the project's principal investigator, Prof. Tirosh-Becker (see examples of her relevant previous scholarship: Tirosh-Becker 1989, 2012, 2015b, 2021).

3.4.1 Digitization

All the texts selected for TAJA were digitized to enable further processing. As JA is written in Hebrew script, one might assume that JA texts, at least printed JA books, could be automatically digitized using Hebrew language OCR (Optical Character Recognition) software. Unfortunately, automated digitization with OCR software failed with these JA texts for two reasons. First, the fonts used in these old books and journals are not identical to those of standard Modern Hebrew; they have JA-specific adaptations (such as additional diacritic points above or below specific characters, e.g., ג׆ or גׇ to denote /ǧ/ vs. ג to denote /ǧ/) and vary from one printing house to another. Second, these books were often stored under less than favorable conditions due to the hardships of immigration; thus, their physical condition (tears, fading, stains, etc.) also disrupts the utility of OCR. Hence, all TAJA texts were manually typed by the project's team of research assistants and subsequently meticulously proofread.

3.4.2 Annotation

Every word in the TAJA corpus was manually annotated for linguistic features on multiple levels using the tag-set describe below. The two main levels of annotation were part-of-speech (POS) tagging and a full morphological analysis of the word. Additional linguistic information was captured in the 'linguistic comments' field. The linguistic annotation was done manually by expert JA linguists (graduate level RAs) under the supervision of Prof. Tirosh-Becker. The accuracy of the manual POS tagging was evaluated using a sample of 3,685 words, quantitatively comparing the tagging accuracy of a junior expert (a graduate level RA) to that of the senior expert, which is considered the 'gold standard.' The measured tagging accuracy of the junior expert was 0.908 (Kessler 2022:54–55), a number which was later used to evaluate the accuracy of the automatic taggers that are developed based on TAJA (Section 5 below).

3.5 *Linguistic Tags Used in the TAJA Corpus*

A complete set of JA-relevant linguistic tags was developed especially for the TAJA corpus to capture in full the specifics of JA morphology. In recent years a uniform tagging system for language corpora was proposed by the Universal

Dependencies (UD) project.[16] These tags, however, were first introduced in 2015, long after the TAJA project was well underway and a large portion of TAJA was already tagged as described below. Following is a detailed description of the TAJA tag-set.

3.5.1 Part-of-speech (POS)

Each word is tagged with a unique part-of-speech (POS) tag. The tags are drawn from a closed list: *noun, verb, particle, proper noun, relative pronoun, adjective, number, personal pronoun, demonstrative, adverb, presentative, quantifier, acronym.* POS tagging was also applied to the embedded Hebrew/Aramaic/French words, which are identified in the TAJA database by suitable tags, as these embedded words are interwoven into the syntactic fabric of JA. In almost all cases these borrowed words were nouns.

3.5.2 Morphology Tags

The morphology of each word was fully analyzed by expert JA linguists. Since in Semitic languages word morphology is closely related to POS, the tagging system that we developed uses different codes for different POS; i.e., a morphological analysis of a *verb* requires a different set of tags than a morphological analysis of a *noun*. The details of all the morphological tags for each POS are listed in Table 3 below. Here are a couple of examples:

– A morphological analysis of a VERB includes *lemma, root, stem*, *tense*, and *person,* and may also include an accusative *pronominal suffix*. For example, the verb ונג'עלך (*u-nəǧʿal-ək* 'and I will appoint you') is analyzed as follows: Lemma: נג'על (*nəǧʿal*), Root: ג'על (*ǧʿl*), Stem: verbal Form I, Tense: imperfect, Person: 1Sg, Pronominal suffix: 2mSg.

Word	**POS**	**Lemma**	**Root**	**Stem**	**Tense/ Mood**	**Person**	**Pron. suffix**
ונג'עלך (*u-nəǧʿal-ək*)	verb	נג'על (*nəǧʿal*)	ג'על (*ǧʿl*)	Form I	Imperfect	1Sg	2mSg

16 See https://universaldependencies.org/ (last accessed May 5, 2022).

– A morphological analysis of a NOUN includes *lemma, gender, number, diminutive*, and *possessive pronominal suffix*. For example, the noun מנאזלך (*mnāzl-ək* 'your dwellings') is analyzed as follows: Lemma: מנאזל (*mnāzəl*), Gender: masculine, Number: plural, Pronominal suffix: 2nd person.

Word	POS	Lemma	Gender	Number	Pron. Suffix
מנאזלך (*mnāzl-ək*)	noun	מנאזל (*mnāzəl*)	m	plural	2nd person

TABLE 3 POS and morphology tags for the TAJA corpus

POS Category	Morphology Tags
Verb	Verbal stem (Form I, Form II, etc.), tense/mood (perfect, imperfect, imperative, active participle, passive participle, etc.), person (1Sg, 2mSg, 2fSg, etc.), accusative pronominal suffix
Noun	Gender (M, F), number (Sg/dual/Pl), diminutive, possessive pronominal suffix
Proper name	Class (person, place, nation, etc.), type (Arabic, Arabicized, Hebrew)
Adjective	Gender (m, f), number (Sg/Pl), comparative form
Adverb	Adverb of time, adverb of place, etc.
Relative pronoun	(none)
Personal pronoun	Person (1Sg, 2mSg, 2fSg, etc.)
Demonstrative pronoun	Gender (m, f), number (Sg/Pl), deixis (near, distant)
Particle	Negation, interrogative, purpose or intent, possession, preposition, conjunction, comparative, conditional, emphasis, contrast, exception and restriction, interjection, disjunction, concession
Presentative	(none)
Adjunct	type (temporal, locative)
Quantifier	(none)
Number	Gender (m, f), cardinal/ordinal
Acronym	(none)
Loanwords	Tags that indicate embedded non-Arabic words or phrases: Hebrew words (including Aramaic), non-Hebrew words (French, Spanish, etc.)

3.6 *Statistics of the TAJA corpus*

The TAJA corpus currently includes over 60,000 words from the 17 Algerian JA texts listed in Table 1 above. Table 4 summarizes the number of tokens and types in TAJA. The portion of the TAJA corpus that was not annotated (approximately 3%) included quotations from the Hebrew Bible and from Hebrew rabbinic literature that appeared in these JA texts. These quotations were not annotated in TAJA because speakers of the dialect do not consider them part of JA (in other words, code switching is easily identified in the context of these Hebrew quotations). In contrast, Hebrew, Aramaic, and other foreign words and phrases that are fully integrated within JA were annotated with the rest of the text, as they are considered integral to JA and not viewed as code switching.

The ratio between the 'number of unique words' (types) and the 'total number of words' (tokens) in the corpus (known as the type-to-token ratio, or TTR) is a measure of the lexical richness of the corpus. The TAJA corpus has a TTR ratio of 0.29, which indicates that from a lexical perspective it is a rich and diverse corpus, especially as most of its texts are long. For comparison, the Penn Arabic Treebank corpus (Maamouri et al. 2004), which is based on newswire texts, has a significantly lower TTR ratio of only 0.13, although this difference may also result from differences in text size.

The distribution of part-of-speech (POS) tags in the TAJA corpus is summarized in Fig. 1. Not surprisingly, the two most prevalent parts of speech are nouns (32.2%) and verbs (26.2%). These are followed by a high prevalence of particles (22.4%) and a relatively low prevalence of adjectives (3.1%).

4 The Utility of the TAJA Corpus for the Study of Algerian JA

The TAJA corpus has already proven instrumental for in-depth analysis of Algerian Judeo-Arabic. The fine-detail level of linguistic annotation of this corpus enables complex queries that shed light on a variety of linguistic aspects within the grammar of Algerian JA. Here we briefly review a few findings

TABLE 4 Size of the annotated TAJA corpus

Corpus	Total words	Annotated		
		Sentences	Words ("Tokens")	Unique words ("Types")
TAJA	63,158	9,904	61,481	17,876

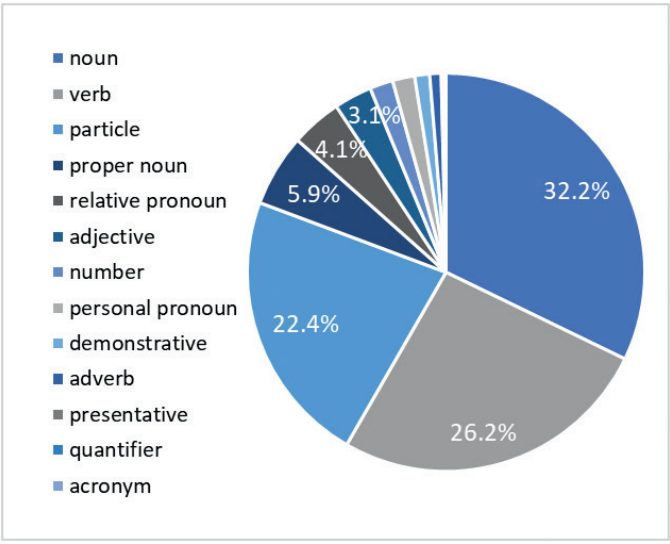

FIGURE 1 Distribution of part-of-speech (POS) tags in the TAJA corpus

that were reported in detail elsewhere: (i) an analysis of dialectal roots in the conservative JA language of *šarḥ* translations (Tirosh-Becker 2011a), and (ii) linguistic characterization of JA registers (Tirosh-Becker 2014). Additional TAJA-based studies elucidating other aspects of the grammar of Algerian JA have been reported at conferences and are expected to be published soon.

4.1 *Dialectal Roots in JA Šarḥ Translations*

The linguistically annotated TAJA corpus is instrumental for analyzing the presence of dialectal roots in the literary genre of the *šarḥ* (pl. *šurūḥ*), i.e., JA translations of biblical books and of post-biblical texts, which is well represented in the TAJA corpus. The language of the *šurūḥ*, in particular that of Bible translations, is uniquely characterized by the significant presence of conservative and archaic linguistic components that are either rare in the spoken dialect or have ceased to exist altogether. Examples of archaic language phenomena preserved in the *šarḥ* to the Bible from Constantine, Algeria, are the use of the n-form *nəktəb* ('it was written') to denote the passive voice of the simple verbal stem (**inCaCaCa* > *nəCCəC*), preservation of the distinct morpheme *–āt* for feminine plural vs. the colloquial use of the masculine plural suffix *–īn* also for plural feminine forms (e.g., *ṭāḥṛāt* vs. *ṭāḥṛīn* 'pure (pl.)'), and the use of the archaic plural demonstrative pronoun *hāwlay* (האולאי 'these') and not the colloquial pronoun *hādu* ('these').

Nonetheless, even in this linguistically conservative literary genre there are some colloquial elements, reflecting a process of slow penetration from

the spoken dialect. Examples of colloquial features present in the *šarḥ* from Constantine are the 2nd person plural suffix *–tīw*, e.g., in the perfect form of the 1st stem *CCəCtīw* (*ktəbtīw* 'you (pl.) wrote'), the colloquial verbal stem *CCāC* (*smān* 'became fat'), and the colloquial reflexive/passive verbal stem *ttəCCəC* (e.g., Tirosh-Becker 1989, 2006, 2012). Penetration of colloquial features into the linguistic fabric of the *šarḥ* translations, which were orally transmitted for generations as cultural heirlooms, is an indication of their contact with colloquial dialects.

Of special interest is the penetration of dialectal roots into a variety of literary JA genres, and in particular into the conservative language of the *šarḥ*. Querying a slightly earlier version of the TAJA corpus, we identified a variety of vernacular roots that found their way into the Algerian *šarḥ* and other Algerian JA texts. The study identified two types of such dialectal roots: (i) secondary roots derived from colloquial nouns; these included the roots *√lss*, *√sgm*, and *√tkl*, and (ii) roots formed through metathesis; these included the roots *√hʿd*, *√ṣnt*, *√wǧb*, and *√nʿl*. This study was made possible by the annotated TAJA corpus, which enabled querying for these roots and forms within multiple different texts. A detailed account of this study was published in Tirosh-Becker (2011).

4.2 *Linguistic Characterization of JA Registers*

A register is a language variety used for a particular purpose or in a particular situation. Authors and speakers use different registers when targeting different audiences, as the various groups do not necessarily share the same language proficiencies or scholarly background. In particular, texts that aim to preserve and provide accessibility to past traditions are often intended for diverse target audiences, which range from scholars to laypeople. By employing different registers, and sometimes altogether different languages, such educational objectives may be met.

A case study that exemplifies a complex use of registers is the book *Sheveṭ Yehuda*. In many Jewish communities it is customary to recite Hebrew penitential poems, known as *piyyuṭe Səliḥot,* in the weeks that precede the Jewish New Year. However, because knowledge of Hebrew in early 20th century Algeria was limited to scholarly circles, these Hebrew poems were often poorly understood by the common community members. The book *Sheveṭ Yehuda* (Constantine 1936) was published by the prominent Algerian Rabbi Yosef Renassia in an attempt to make the content of these poems accessible to as many members of his community as possible. To that end he bound together in a single book four related texts: (i) a JA translation of 16 Hebrew *Səliḥot,* (ii) a JA commentary on these 16 translated poems, (iii) a French translation for 6 of the Hebrew *Səliḥot*, and (iv) a textual JA guide to the rites and customs of the Jewish New Year. The

author clearly expected the JA texts to be better understood by his community members than the original Hebrew texts. However, recognizing that JA was gradually losing ground in favor of French, he also included a French translation of the most popular *Səliḥot* to help those who no longer commanded JA, let alone Hebrew.

The three JA texts in this book are interesting as they reflect different registers: a *translation* (of the Hebrew text), a *commentary* (on the Hebrew text), and a *freely written* text (explaining the rites and customs of the holiday). As this book is part of TAJA, and since all its JA texts were linguistically annotated in TAJA, we were able to use the corpus to explore the language variations employed in each register.

The analysis showed that the register used for the *translation* of these poems tries to imitate the conservative language of the biblical *šarḥ*, and its register includes some of its characteristics. This is most likely because the *Səliḥot* poems are an integral part of the revered Jewish liturgical corpus. The *commentary*, on the other hand, was written in the literary register employed in the didactic texts published by the community's rabbinic elite for its own use. Finally, the *freely written* text employed a more vernacular register than the other two as it targeted the broadest audience and aimed to be understood by most members of the community, to ensure their proper preparation for the coming holidays. An example of quantitative data from TAJA that support this conclusion is the prevalence of Arabic, Hebrew, and French words in the three texts. While the *translation* is written almost purely in Arabic (99% of the text) almost without any Hebrew elements (1%) and with no French elements at all, the other two texts employ different registers, with many Hebrew loanwords, and even a limited presence of French (0.2% in both texts). As summarized in Table 5 below, the extent of Hebrew loanwords is somewhat greater in the *freely written* text compared to the text of the *commentary* (77% Arabic and 23% Hebrew vs. 67% Arabic and 33% Hebrew in the *commentary* vs. the *freely written* text, respectively). More data about these registers was published in Tirosh-Becker (2014).

TABLE 5 Prevalence of Arabic, Hebrew, and French words in the three JA texts of *Sheveṭ Yehuda*

Register	Arabic	Hebrew/Aramaic	French
Translation text	99%	1%	None
Commentary text	77%	23%	0.2%
Freely written text	67%	33%	0.2%

5 Automated Machine Learning Tools Based on TAJA

The linguistically annotated TAJA corpus has already proven valuable for the study of Algerian JA. The next step, however, is to use the annotated TAJA corpus as a training dataset for developing Natural Language Processing (NLP) machine learning tools that will allow us to expand the linguistic study of Algerian JA to additional *unannotated* texts, and possibly to other North African JA varieties as well. This project, which is already well underway, is done in collaboration with Dr. Yonatan Belinkov and his research group,[17] developing machine learning tools based on the annotated TAJA corpus. Two assumptions underlie this work. The first is that machine learning methodologies that have been successfully developed for the field of NLP are suitable for linguistic analysis of large textual corpora. This premise is already widely accepted and validated (e.g., Nivre 2008, van den Bosch 2009). The second assumption is that the manually annotated TAJA corpus, described herein, is a robust foundation on which machine learning for written Algerian JA can be based.

Two machine learning tools have already been developed based on TAJA (Kessler 2022). These tools are as follows:

1. A *POS tagger* that is capable of assigning part-of-speech tags to JA words.
2. An automated *morphology tagger* that identifies the morphology of the JA words in untagged texts based on the pre-determined (manually or automatically) assignment of POS tags. Namely, when assigning morphological tags (person, stem, tense, etc.) to new words, the machine takes into account the POS information about that word (i.e., whether the word is a noun, a verb, etc.).

The performance of the morphology tagger is better than 90% accuracy, and in some cases, such as enclitics tags, the accuracy is greater than 96%. For a detailed technical description of the models, the TAJA-based training process, and the performance evaluation measures, see Kessler (2022).

6 Conclusions

The Tagged Algerian Judeo-Arabic (TAJA) corpus is the first linguistically annotated (part-of-speech and morphology) corpus of written Algerian Judeo-Arabic, and more generally of any Judeo-Arabic dialect regardless of geography and time. As such, it has already demonstrated its value by preserving important Algerian JA texts from the late 19th and early 20th centuries and

17 The Department of Computer Science, Technion, Israel (formerly of the Departments of Computer Science at Harvard and at MIT, MA).

supporting the linguistic characterization of this language variety. However, the importance of the TAJA corpus goes beyond the study of Algerian Judeo-Arabic *per se*. It pioneers the introduction of Natural Language Processing (NLP) and Machine Learning (ML) methodologies into the field of Judeo-Arabic. First and foremost, the TAJA corpus has already allowed us to develop automated POS and morphology taggers specifically tailored for Algerian JA. The availability of these new NLP taggers will enable us to significantly broaden the scope of our research to the vast textual collections written in Algerian JA—whose analysis is beyond the capacity of any individual scholar. In fact, we are now constructing an additional corpus (NAJA = New Algerian Judeo-Arabic corpus) of untagged JA texts to which we plan to apply these automatic taggers.

Furthermore, we believe that NLP tools also provide new *quantitative* measures that will enable the application of new quantitative approaches to the field of Jewish languages, offering new ways to compare literary genres, language registers, and possibly even neighboring varieties. For example, our preliminary data suggests that a quantitative assessment of literary genre could be calculated by training the tagger on texts from one genre and then applying it to texts from a different genre that were not part of the training set. The accuracy of such tagging would reflect the level of linguistic similarity between the two genres. We believe that the work presented in this article is only the first step towards a more extensive introduction of NLP and ML to the study of Judeo-Arabic and of Jewish languages.

Textual Sources for the TAJA Corpus

#	Text	Reference
1	JA translation of Psalms (chapters 42–50)	Renassia, Yosef. 1954(?). *Zikhron Yaʿakov*, 5 vols. Djerba: Ḥadad imprimerie.
2	JA translation of Proverbs	Renassia, Yosef. 1916. *Azharot Ben David*. Tunis.
3	JA translation of Joshua (Chapters 1–4)	Ha-Cohen, David, Shelomo Zerbib & Zion Shuqrun. 1911. *Sefer Divre Ḥakhamim*. Tunis.
4	JA translation of the Hafṭarot (for Genesis and Exodus)	Renassia, Yosef. (after) 1934. *Sefer Pəṭirat Moshe*. Constantine.
5	JA translation of Passover Haggadah	Renassia, Yosef. 1962. *Zeved Ṭov*, Djerba. reprinted Jerusalem 1986.

(*cont.*)

#	Text	Reference
6	JA translation of the Hosha'anot	Renassia, Yosef. 1930. *Kibbud ʾAv va-ʾEm*. Djerba. reprinted Jerusalem 1987.
7	JA translation of the Seliḥot	Renassia, Yosef. 1933(?). *Sefer Yeme Ḥaninah ve-Sheveṭ Yehudah*. Constantine.
8	JA translation of *Mishnah ʾAvot*	Renassia, Yosef. 1916. *Sefer Milḥamah be-Shalom*. Tunis.
9	JA translation of Maimonides' *Mishne-Torah* (chapter on *Hilkhot Sekhirut*)	Renassia, Yosef. 1954. *Code Israélien civil et religieux—Yad H'azak'ah.* Djerba.
10	JA commentary on the Hosha'anot	Included with no. 6 above.
11	JA commentary on Joshua	Included with no. 3 above.
12	JA commentary on *Mishnah ʾAvot*	Included with no. 8 above.
13	JA commentary on the Seliḥot	Included with no. 7 above.
14	The book *Peraḥ Shoshan* by Shalom Bekache	Bekache, Shalom. 1892. *Sefer Peraḥ Shoshan.* Algeria: R. Shalom Bekache Print.
15	The book *ʾOr ʿOlam* by Shalom Bekache	Included with No. 14 above.
16	JA journal *əl-Ḥikma*	*əl-Ḥikma, Journal littéraire hebdomadaire*, editor: Avraham Zerbib, Constantine, issue no. 6, 1912.
17	JA journal *Maguid Micharim*	*Maguid Micharim, Journal Hebreu-Arabe*, editor: Elie Karsenty, Oran, issue no. 44, 1896.

References

Abidi, Karima, Mohamed Amine Menacer, & Kamel Smaili. 2017. "CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube." 18th Annual Conference of the International Communication Association (Interspeech), Stockholm, Sweden.

Ahmed, Mohamed A. H. 2018. "XML Annotation of Hebrew Elements in Judeo-Arabic Texts." *Journal of Jewish Languages* 6.2: 221–242.

Ahmed, Mohamed A. H. 2022. "18th-Century Judeo-Arabic Documents from the Prize Papers Collection." *Journal of Jewish Languages*, 10.1: 1–23.

Alansary, Sameh, Magdy Nagi, & Noha Adly. 2007. "Building an International Corpus of Arabic (ICA): Progress of Compilation Stage." In *Proceedings of the 7th International Conference on Language Engineering*, Cairo.

Almeman, Khalid & Mark Lee. 2013. "Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words." 1st ICCSPA Conference, Sharjah, 1–6.

Alrabiah, Maha, AbdulMalik Al-Salman, & Eric Atwell. 2013. "The Design and Construction of the 50 Million Words KSUCCA." In *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics*, Lancaster, The University of Leeds, 5–8.

Al-Thubaity, Abdulmohsen O. 2015. "A 700M+ Arabic Corpus: KACST Arabic Corpus Design and Construction." *Language Resources and Evaluation*, 49(3): 721–751.

Arts, Tressy, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff, & Vit Suchomel. 2014. "arTenTen: Arabic Corpus and Word Sketches." *Journal of King Saud University—Computer and Information Sciences* 26.4: 357–371.

Bar, Kfir, Nachum Dershowitz, Lior Wolf, Yackov Lubarsky, & Yackov Choueka. 2015. "Processing Judeo-Arabic Texts." In *Proceedings of the First International Conference on Arabic Computational Linguistics* (*ACLing*), Cairo 2015, 138–144.

Bar-Asher, Moshe. 1992. *La composante hebraïque du judeo-arabe Algerien: communautes de Tlemcen et Aïn-Temouchent*. Jerusalem: Magnes.

Bar-Asher, Moshe. 1999. *Traditions Linguistiques des Juifs d'Afrique du Nord*, 2nd edition. Jerusalem: The Hebrew University, Section 1, 3–129 (in Hebrew).

Bar-Asher, Moshe. 2002. *Le Commentaire biblique Leshon limmudim de Rabbi Raphaël Berdugo*. Jerusalem: The Hebrew University (in Hebrew).

Bar-Asher, Moshe. 2005. "The Judeo-Arabic of Tunisia." In *Tunisia, Jewish Communities in the East in the Nineteenth and Twentieth Centuries*, ed. Haim Saadoun. Jerusalem: The Ben-Zvi Institute, 269–274 (in Hebrew).

Bar-Asher, Moshe. 2017. "Edited Documents from Ghardaia." In *Ḥiqrey Maʿarav: Studies in the Languages, Traditions, Customs, and Documents of the Maghrebian Jews.* New Haven: Yale University, Section 4 (chapters 14–16), 277–318 (in Hebrew).

Belinkov, Yonatan, Alexander Magidow, Alberto Barrón-Cedeño, Avi Shmidman, & Maxim Romanov. 2019. "Studying the History of the Arabic Language: Language Technology and a Large-Scale Historical Corpus." *Language Resources and Evaluation* 53: 771–805.

Boucherit, Aziza. 2002. *L'arabe parlé à Alger: Aspects sociolinguistiques et énonciatifs.* Paris-Louvain: Peeters.

Bougrine, Soumia, Aicha Chorana, Abdallah Lakhdari, & Hadda Cherroun. 2017. "Toward a Web-based Speech Corpus for Algerian Arabic Dialectal Varieties," In *Proceedings of The Third Arabic Natural Language Processing Workshop* (*WANLP*). Valencia, Spain: Association for Computational Linguistics, 138–146.

Brunot, Louis & Élie Malka. 1939. *Textes judéo-arabes de Fès*. Rabat: Typo-litho École du livre.

Brunot, Louis & Élie Malka. 1940. *Glossaire judéo-arabe de Fès*. Rabat: Typo-litho École du livre.

Bunis, David Monson. 1993. *A Lexicon of the Hebrew and Aramaic Elements in Modern Judezmo*. Jerusalem: Magnes Press.

Cantineau, Jean. 1936. "Géographie linguistique des parlers arabes algériens." *Revue Africaine* 79: 91–93.

Cantineau, Jean. 1937. "Les parlers arabes du département d'Alger." *Revue Africaine* 81: 703–711.

Cantineau, Jean. 1938. "Les parlers arabes du département de Constantine." *IVe Congrès de la Fédération des Sociétés savantes de l'Afrique du Nord*, 2, 849–863.

Cantineau, Jean. 1940. "Les parlers arabes du département d'Oran." *Revue Africaine* 84: 220–231.

Cantineau, Jean. 1941. "Les parlers arabes des territoires du sud." *Revue Africaine* 85: 72–77.

Caubet, Dominique. 1993. *L'arabe marocain*. Paris: Peeters.

Chetrit, Joseph. 2010. *Trésors et textures d'une langue: études socio-pragmatiques sur le judéo-arabe en Afrique du Nord et son composant hébraique—articles, poèmes, récits et proverbs*. Jerusalem: Bialik Institute (in Hebrew).

Chetrit, Joseph. 2016. "Diversity of Judeo-Arabic Dialects in North Africa: Eqaːl, Wqal, kʲal and ʔal Dialects." *Journal of Jewish Languages* 4.1: 1–43.

Cohen, David. 1970. "Les deux parlers arabes de Tunis: Notes de phonologie comparée." In *Études de linguistique sémitique et arabe.* The Hague: Mouton, 150–171.

Cohen, David. 1975. *Le parler arabe des Juifs de Tunis*, vol. 2: *Étude Linguistique*. The Hague: Mouton.

Cohen, Marcel. 1912. *Le parler arabe des Juifs d'Alger*. Paris: H. Champion.

Cotterell, Ryan & Chris Callison-Burch. 2014. "A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic." *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, 241–245.

D'Anna, Luca. 2021. "The Judeo-Arabic Dialect of Yefren (Libya): Phonological and Morphological Notes." *Journal of Jewish Languages* 9.1: 1–31.

Fischer, Wolfdietrich & Otto Jastrow. 1980. *Handbuch der Arabischen Dialekte*. Wiesbaden: Harrassowitz.

Fleischer, Heinrich L. 1864. "Jüdisch-Arabisches aus Magreb." *Zeitschrift der Deutschen Morgenländischen Gesellschaft* (*ZDMG*) 18: 329–340.

Grand'Henry, Jacques. 1972. *Le parler arabe de Cherchell* (*Algérie*), Louvain-la-Neuve: Université Catholique de Louvain, Institut orientalist.

Habash, Nizar, Owen Rambow, & Ryan Roth. 2009. "MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization." In *Proceedings of the International Conference on Arabic Language Resources and Tools*, Cairo, 102–109.

Hajič, Jan, Otakar Smrž, Petr Zemánek, Jan Šnaidauf, & Emanuel Beška. 2004. "Prague Arabic Dependency Treebank: Development in Data and Tools." *NEMLAR International Conference on Arabic Language Resources and Tools*, Cairo, 110–117.

Harrat, Salima, Karima Meftouh, Mourad Abbas, Salma Jamoussi, Motaz Saad, & Kamel Smaili. 2015. "Cross-Dialectal Arabic Processing." International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, 620–632. https://doi.org/10.1007/978–3-319–18111–0_47.

Heath, Jeffrey & Moshe Bar-Asher. 1982. "A Judeo-Arabic Dialect of Tafilalet (Southeast Morocco)." *Zeitschrift für Arabische Linguistik* 9: 32–78.

Heath, Jeffrey. 1989. *From Code-switching to Borrowing: Foreign and Diglossic Mixing in Moroccan Arabic*. London: Kegan Paul International.

Heath, Jeffrey. 2002. *Jewish and Muslim Dialects of Moroccan Arabic*. London: Routledge Curzon.

Henshke, Yehudit. 2007. *Lashon Ivri bedibur Aravi*. Jerusalem: Bialik Institute (in Hebrew).

Kessler, Michal 2022. *Morphosyntactic Tagging of Algerian Judeo-Arabic*. MS Thesis, School of Computer Science and Engineering, The Hebrew University of Jerusalem.

Khalifa, Salam Magdi, Nizar Habash, Dana Abdulrahim, & Sara Hassan. 2016. "A Large Scale Corpus of Gulf Arabic." In *Proceedings of the Language Resources and Evaluation Conference* (*LREC*), Portorož, 4282–4289.

Laraba, Ahmed. 1981. *A Linguistic Description of the Algerian Arabic Dialect of Constantine*, Ph.D. thesis, Manchester.

Levy, Simon. 2009. *Parlers arabes des Juifs du Maroc: Histoire, sociolinguistique et geographie dialectale*. Zaragoza: Universidad de Zaragoza.

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, & Wigdan Mekki. 2004. "The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus." NEMLAR International Conference on Arabic Language Resources and Tools, Cairo.

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow, & Dalila Tabessi. 2006. "Developing and Using a Pilot Dialectal Arabic Treebank." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, 443–448.

Maamouri, Mohamed, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, & Ramy Eskander. 2014. "Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development." *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, 2348–2354.

Maman, Aharon. 2000. "The Maghrebi Sharḥ of the Bible." *Pe'amim* 83: 48–56 (in Hebrew).

Maman, Aharon. 2019. *Synoptic Dictionary of the Hebrew Component in Jewish Languages: Including the Notes of Shelomo Morag*. Second revised version. Jerusalem: Magnes & The Hebrew University (in Hebrew).

Mangion, M. 1937. "Le dialect arabe de l'Edough." *Revue Africaine* 81: 373–380.

Marçais, Philippe. 1936. "Remarque sur un fait syntaxique du parler arabe d'El-Milia." *Revue Africaine* 79: 1047–1055.

Marçais, Philippe. 1947. "Contribution à l'étude du parler arabe de Bou-Saâda." *Bulletin de l'Institut Français D'archéologie Orientale* 44: 21–88.

Marçais, Philippe. 1954. *Textes arabes de Djidjelli*. Paris.

Marçais, Philippe. 1956. *Le parler arabe de Djidjelli* (*Nord constantinois, Algerie*), Paris: Adrien-Maisonneuve.

Marçais, Philippe. 1977. *Esquisse grammaticale de l'arabe maghrébin*. Paris: Librairie d'Amérique et d'Orient.

Marçais, William. 1902. *Le dialecte arabe parlé à Tlemcen*. Paris: E. Leroux.

Marçais, William. 1908. *Le dialecte arabe des Ūlâd Brāhîm de Säida*. Paris: Honoré Champion.

Matras, Yaron. 2009. *Language Contact*. Cambridge: Cambridge University Press.

Meftouh, Karima, Salima Harrat, Salma Jamoussi, Mourad Abbas, & Kamel Smaili. 2015. "Machine Translations Experiments on PADIC: A Parallel Arabic Dialect Corpus." In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, 26–34.

Nivre, Joakim. 2008. "Treebanks." In *Corpus Linguistics: An International Handbook,* eds. Anke Lüdeling & Merja Kytö. Berlin: Walter de Gruyter, vol. 1, 225–241.

Ostoya-Delmas, S. 1938. "Notes preliminaires a l'étude des parlers de l'arrondissement de Philippeville," *Revue Africaine* 82: 60–83.

Poplack, Shana. 1980. "'Sometimes I'll Start a Sentence in Spanish y termino en Espanol:' Toward a Typology of Code-switching." *Linguistics* 18(7–8): 581–618.

Romanov, Maxim & Masoumeh Seydi. 2019. "OpenITI: A Machine-Readable Corpus of Islamicate Texts (2019.1.1)" [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3082464 (last accessed December 25, 2021).

Saad, Motaz K. & Wesam Ashour. 2010. "OSAC: Open Source Arabic Corpora." In *6th International Conference on Electrical and Computer Systems*, Cyprus, 118–123.

Saada, Lucienne. 1956. "Introduction á l'étude du parler des Juifs de Sousse." *Les Cahiers du Tunisie* 16: 518–532.

Seddah, Djamé, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, & Abhishek Srivastava. 2020. "Building a User-generated Content North-African Arabizi Treebank: Tackling Hell." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, 1139–1150.

Shoufan, Abdulhadi & Sumaya Al-Ameri. 2015. "Natural Language Processing for Dialectical Arabic: A Survey." *Proceedings of the Second Workshop on Arabic Natural Language Processing*, Beijing, 36–48.

Singer, Hans-Rudolf. 1984. *Grammatik der arabischen Mundart der Medina von Tunis*. Berlin: Walter de Gruyter.

Stillman, Norman A. 1988. *The Language and Culture of the Jews of Sefrou, Morocco*. Manchester: University of Manchester Press.

Stumme, Hans. 1896. *Grammatik des Tunisischen Arabisch*. Leipzig: J.C. Hinrichs.

Talmoudi, Fathi. 1980. *The Arabic Dialect of Sūsa* (*Tunisia*). Göteborg: Acta Universitatis Gothoburgensis.

Tedghi, Joseph. 2003. "Évolution des recherches sur la composante hébraïque dans les parlers judéo-arabes maghrébins modernes." In *Linguistique des langues juives et linguistique Générale*, eds. Frank Alvarez-Péreyre & Jean Baumgarten. Paris: CNRS Editions, 157–190.

Tedghi, Joseph. 2012. "'Le livre de Jonas' traduit en judéo-arabe marocain par Samuel Malka: étude linguistique." In *Dynamiques langagières en Arabophonies*, eds. Alexandrine Barontini, Christophe Pereira, Ángeles Vicente, & Karima Ziamari. Zaragoza: Universidad de Zaragoza, 253–290.

Tedghi, Joseph. 2016. "Tunisian Judeo-Arabic." In *Encyclopedia of Jews in the Islamic World*, ed. Norman A. Stillman. Leiden: Brill.

Tirosh-Becker, Ofra. 1989. "A Characterization of the Judeo-Arabic Language of Constantine." *Massorot* 3–4: 285–312 (in Hebrew).

Tirosh-Becker, Ofra. 2006. "An Algerian Judeo-Arabic Translation of Piyyuṭ Mi Khamokha by Rabbi Yehuda Ha-Levi." *Massorot* 13–14: 315–369 (in Hebrew).

Tirosh-Becker, Ofra. 2011a. "On Dialectal Roots in Judeo-Arabic Texts from Constantine (East Algeria)." *Revue des Études Juives* 170.1–2: 227–253.

Tirosh-Becker, Ofra. 2011b. "Terms for Realia in an Algerian Judeo-Arabic Translation of the Hošaʿnot." In *Studies in the Culture of North African Jewry*, eds. Moshe Bar-Asher & Steven D. Fraade. New Haven & Jerusalem: Yale University & The Hebrew University, vol. 1, 171–186.

Tirosh-Becker, Ofra. 2011c. "Archaic and Dialectal Features in an Algerian Judeo-Arabic Translation and Commentary of Tractate Avot." In *Hikrei Ma'arav u-Mizraḥ: Studies in Language, Literature and History Presented to Joseph Chetrit*, eds. Yosef Tobi & Dennis Kurzon. Jerusalem: Carmel, 181–207 (in Hebrew).

Tirosh-Becker, Ofra. 2012. "Mixed Linguistic Features in a Judeo-Arabic Text from Algeria: The *Šarḥ* to the *Haftarot* from Constantine." In *Language and Nature: Papers Presented to John Huehnergard on the Occasion of his 60th Birthday*, eds. Rebecca Hasselbach & Na'ama Pat-El. Chicago: The Oriental Institute, 391–406.

Tirosh-Becker, Ofra. 2013. "Algeria." In *Encyclopedia of Hebrew Language and Linguistics*, ed. Geoffrey Khan. Leiden: Brill, vol.1, 85–86.

Tirosh-Becker, Ofra. 2014. "A Reflection of a Linguistic Reality: An Algerian Judeo-Arabic Book for the New Year." In *Studies in the Culture of North African Jewry,* eds. Moshe Bar-Asher & Steven D. Fraade. New Haven & Jerusalem: Yale University & The Hebrew University, vol. 3, 193–216.

Tirosh-Becker, Ofra. 2015a. "Eli'ezer Ben-Yehuda and Algerian Jews: Relationship and Language." In *Arabic and Semitic Linguistics Contextualized. A Festschrift for Jan Retsö*, ed. Lutz Edzard. Wiesbaden: Harrassowitz Verlag, 430–447.

Tirosh-Becker, Ofra. 2015b. "Two Judeo-Arabic Translations of the Scroll of Antiochus from Ghardaia (Algeria)." In *Darchei Noam: The Jews of Arab Lands*, eds. Carsten Schapkow, Shmuel Shepkaru, & Alan T. Levenson. Leiden: Brill, 185–213.

Tirosh-Becker, Ofra. 2017. "Hebrew and Judeo-Arabic in Homilies for Bar Mitzva from Ghardaia (Algeria)." *Language Studies* 17–18: 611–636 (in Hebrew).

Tirosh-Becker, Ofra. 2019. "Linguistic Analysis of an Algerian Judeo-Arabic Text from the 19th Century." *La Linguistique* 55.1: 192–211.

Tirosh-Becker, Ofra. 2021. "North African Judeo-Arabic." In *Jewish Languages: Text Specimens, Grammatical, Lexical, and Cultural Sketches*, eds. Lutz Edzard & Ofra Tirosh-Becker. Porta Linguarum Orientalium. Wiesbaden: Harrassowitz Publishers, 252–294.

van den Bosch, Antal. 2009. "Machine Learning." In *Corpus Linguistics: An International Handbook*, eds. Anke Lüdeling & Merja Kytö. Berlin: Walter de Gruyter, vol. 2, 855–874.

Yoda, Sumikazu. 2005. *The Arabic Dialect of the Jews in Tripoli* (*Libya*)*: Grammar, Text and Glossary*. Wiesbaden: Harrassowitz Verlag.

Yoda, Sumikazu. 2006. "'Sifflant' and 'Chuintant' in the Arabic Dialect of the Jews of Gabes (south Tunisia)." *Zeitschrift für Arabische Linguistik* 46: 7–25.

Yoda, Sumikazu. 2010. "Libyan Judeo-Arabic." In *Encyclopedia of Jews in the Islamic World*, ed. Norman A. Stillman. Leiden: Brill.

Zafrani, Haim. 1967. "Les langues juives du Maroc." *Revue de l'occident et de la méditerranée* 4: 175–188.

Zaghouani, Wajdi. 2014. "Critical Survey of the Freely Available Arabic Corpora." In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*, Reykjavik, 1–8.

Zaidan, Omar F. & Chris Callison-Burch. 2011. "The Arabic Online Commentary Dataset: An Annotated Dataset of Informal Arabic with High Dialectal Content." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*. Portland, Oregon, 37–41.

Ofra Tirosh-Becker

is a professor, Department of Arabic Language and Literature, and the Department of Hebrew Language, The Hebrew University of Jerusalem, Israel. She is the Head of its Center for Jewish Languages and Literatures. Her publications include *Rabbinic Excerpts in Medieval Karaite Literature* (Jerusalem 2011), and *Jewish Languages: Text Specimens, Grammatical, Lexical, and Cultural Sketches* (with Prof. Lutz Edzard; Wiesbaden 2021).

Oren M. Becker
is an expert in bioinformatics and computational biophysics. He is a serial entrepreneur, executive and board member of innovative biopharmaceutical companies, and the inventor of several computational drug discovery technologies. Oren M. Becker is a former professor of computational biophysics, Department of Chemical Physics, Tel Aviv University, Israel, and a visiting professor at Harvard University.