Towards a Computational Lexicon for Moroccan Darija: Words, Idioms, and Constructions

Jamal Laoudi[⋆], Claire Bonial[†], Lucia Donatelli[‡], Stephen Tratz[†], Clare Voss[†]
[⋆]ARTI, Alexandria, Virginia 22030
[†]Army Research Laboratory, Adelphi, Maryland 20783
[‡]Georgetown University, Washington, DC 20057
{jamal.laoudi.ctr, claire.n.bonial.civ, stephen.c.tratz.civ,
clare.r.voss.civ}@mail.mil, led66@georgetown.edu

Abstract

We explore the challenges of building a computational lexicon for Moroccan Darija (MD), an Arabic dialect spoken by over 32 million people worldwide that only recently has begun appearing frequently in written form. We raise the question of what belongs in such a lexicon and start by describing our work building traditional word-level lexicon entries with their English translations. We then discuss challenges in translating idiomatic MD phrases and the creation of multi-word expression (MWE) lexicon entries whose meanings could not be fully derived from the individual words. Finally, we describe our preliminary exploration of constructions for inclusion in an MD *constructicon*, initially eliciting translations of established English constructions, and then shifting to document, when spontaneously offered, variant renderings of native MD counterparts.

1 Introduction

What methods exist to guide the construction of a computational lexicon for a low-resource language that is widely spoken, but for which there is little reference literature and no standard orthography? We are interested in the specific case of Moroccan Darija (MD) that is now emerging in the written, informal contexts of social media. There, we find a wide variety of multi-word expressions (MWEs), including idioms, that are characterized by semi- or non-compositionality, where their meaning cannot be derived strictly from their individual words. Such expressions therefore present a significant challenge for Natural Language Processing (NLP), language-learning, and machine translation (MT) tasks, which have traditionally assumed that the vocabulary of the relevant language is available in a (computational) lexicon so that a sentence's full meaning can be derived from the combined meanings of its individual words. For MT of under-resourced languages, and MD in particular, available methods assume that translation proceeds from the source language (SL) by way of entries in a limited bilingual dictionary, into the corresponding target language (TL). In the case of semi- and non-compositional expressions of low-resource source languages, however, well-established methods do not exist for identification and inclusion of such expressions in computational lexicons.

As a result, correctly interpreting and translating semi- and non-compositional expressions relies on first identifying MWEs that function as a unit paired with a particular meaning. Identification can be facilitated by manually tagging such expressions and incorporating them into lexical resources. However, incorporation into a lexicon can also come with a variety of challenges, as outlined by Sag et al. (2002). Fixed idiomatic expressions, such as *kick the bucket* can be added to a lexicon as a single-entry, fixed phrase allowing only for minimal morpho-syntactic variation (here, tense), so that the phrase is treated computationally as if it were a single word (an approach called 'words with spaces' (Sag et al., 2002)). However, many other flexible expressions allow for enough variation that listing all alternates in a computational lexicon can be prohibitively impractical. Similarly, fully syntactic patterns can carry meaning separate from the constituent parts and be extremely productive, or flexible, in the component words (e.g., *The X-er The Y-er: The more I read, the less I understand; The higher you fly, the harder you fall*). Such pairings of form and meaning, or constructions (Goldberg, 1995), are productive to the

Proceedings of the Joint Workshop on
Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 74–85
Santa Fe, New Mexico, USA, August 25-26, 2018.

point that listing all possible instantiations of the expression in a static lexicon is not feasible. As an alternative, these patterns can be listed in a *constructicon*, or an inventory of constructions in which the general linguistic patterns are described without specifying all realizations with particular word forms (Fillmore et al., 2012).

In this research, we begin to address the challenges of semi- and non-compositional MWEs in MD that arise while building up a MD lexicon consisting of 1) individual MD tokens and their translations, 2) a collection of MD idioms, and 3) a preliminary constructicon listing MD constructions based on an examination of constructions found in English and their correspondents in MD. Our objective in building the MD lexicon is to facilitate translation of MD into English and vice-versa. After providing some background on MD (Section 2), we will describe each portion (1-3 above) of the lexicon in Sections 3-5. We examine translation in the order of words, idioms, then constructions as this reflects compatibility with traditional lexicons: words are traditional members of lexicons, some (more fixed) idioms are compatible with a 'words-with-spaces' lexicon entry, and constructions are the least compatible in that they require an inventory that represents grammatical patterns as opposed to fixed word forms or expressions. Of current interest is the contribution of a collection of MD idioms; thus, in Section 4 describing the MD idioms, we provide an analysis of the strategies used to translate idioms as well as a quantitative results illuminating which strategies were applicable to particular idiom types.

2 Moroccan Darija: an Arabic Dialect

The Arabic language family encompasses many varieties, including Modern Standard Arabic (MSA) and a large number of regional dialects. Despite being the standard written form of Arabic and appearing in many formal venues such as news broadcasts and parliamentary speeches, MSA is not the native language of any Arabic speaker (Bouamor et al., 2014). Instead, typical everyday conversation is conducted in Arabic dialects (or other languages) rather than MSA. Arabic dialects vary substantially from MSA and each other, with significant phonological, morphological, and syntactic differences (Brustad, 2002), as well as in their vocabularies. For example, MSA has been established as VSO, whereas MD is viewed as SVO (Simons and Fennig, 2018). The dialects have varying degrees of mutual intelligibility, with those of northwestern Africa (e.g., Morocco, Algeria, Tunisia) being particularly challenging to outsiders.

Although Arabic dialects lack orthographic standards and historically have not been written, they now appear constantly in online social media. Unfortunately, these dialects are severely under-resourced from a computational perspective, impeding the creation of NLP systems, such as machine translation engines. MT systems trained on large parallel MSA-English corpora perform substantially worse on dialectal text than when such systems are trained on significantly smaller Arabic dialect-English parallel corpora created via crowdsourcing (Zbib et al., 2012), illustrating the need for computational lexical resources specific to the individual dialects.

3 Traditional Lexicon: Words and Their Translations

The most straightforward component to include in an MD computational lexicon is a traditional word-level lexicon consisting of words and their translations. For our work, we leverage Hespress.com, a popular news portal whose reporting covers events in Morocco and other topics of interest to Moroccans. The articles span numerous domains, including politics, sports, and entertainment. Although the articles are primarily written in MSA, the commentary is mixed, with many languages/dialects represented, including MSA, MD, English, French, and Amazigh. As a first step, we collected a copy of the commentary available on the Hespress website and filtered out commentary posts that lacked content in Arabic script.[1] To isolate posts written primarily in MD, we employed an in-house token-level MD classifier[2] to label the individual tokens[3] of the remaining posts as either MD or MSA. We only considered commentary posts with at least 6 tokens identified as MD and with MD tokens outnumbering MSA tokens by a 4-to-1 or greater margin. This left us over 128,000 commentary posts identified as primarily MD. We then

[1] Arabic dialects such as MD also frequently appear in Latin script as "Arabizi" (Yaghan, 2008).

[2] The full details of this system are beyond the scope of this paper.

[3] We tokenized the text automatically based on whitespace and punctuation.

had a native Darija speaker translate the 2,000 shortest such commentary posts and provide in-context token-level glosses. After removing posts written in other languages (e.g., MSA, Amazigh) or containing offensive content, 1,836 translated posts remained, with a total of 17,261 tokens and 5,528 unique token types. The 5,528 types have an average of 2.32 glosses per type.

The only machine-readable lexicon for MD we are aware of currently available for comparison is published by the MADAR (Bouamor et al., 2018) project. The MADAR lexicon is structured around 1,045 *concept keys*, each of which is defined by a 3-tuple of English, French, and Modern Standard Arabic (MSA) lexical translations (e.g., {*think, penser,* ظَنَّ }). For each concept key, MADAR had Arabic dialect translations produced by translators in 25 different cities spanning the Arabic-speaking world, including two Moroccan cities—Fes and Rabat.

We note substantial overlap with our lexicon. Of the 1,032 unique MD translations provided by MADAR's Fes and Rabat translators, 342 correspond to items in our lexicon. However, there are several important differences between the two lexicons. For example, the MD entries in MADAR were all generated in translation from other languages whereas ours were originally produced in MD and then translated into English. Because of this, our lexicon is likely to contain a variety of morphological and orthographic variations for the same base word, in contrast to the MADAR lexicon. Also, the MADAR concept keys were derived from the Basic Travel Expressions Corpus (Takezawa et al., 2007) and, thus, the lexicon is most relevant to the travel domain. In contrast, our lexicon spans multiple complementary domains, including politics, sports, and entertainment.

4 More than a Word: Idioms

In the process of creating the word-level lexicon described in Section 3, we encountered numerous idioms. Idioms pose an intriguing question in the development of a lexicon, as their status straddles the boundary between being compositional and non-compositional in meaning. A central challenge in identifying and translating idioms is thus pinpointing this variable idiomatic meaning and transferring the semantic, stylistic, and contextual import of the idiom in question from SL to TL. Here we review several theories for interpreting idiomatic meaning. We then present the strategies for idiom translation that our translator found consistently helpful in guiding his work in reference to other established idiom translation techniques.

4.1 Idiomatic Meaning

Nunberg et al. (1994) analyze idioms for three primary semantic properties: (i) *conventionality*, measured by the discrepancy between idiomatic phrasal meaning and the compositional meaning derived from the individual constituents; (ii) *opacity/transparency*, or the clarity of motivation for the expression given its meaning; and (iii) *compositionality*, understood as the degree to which the phrasal meaning can be gleaned from the component constituents. Fernando (1996) additionally divides idioms into three classes, each with distinct measures of these semantic properties and, in turn, distinct implications for translation. So-called "pure idioms" are opaque, non-literal MWEs that are conventionalized in use and non-compositional (cf. Sag et al. (2002) for non-decomposable idioms and their syntactic diagnostics); "semi-idioms" are semi-opaque and have one or more literal constituents and one with a non-literal subsense; and "literal idioms" are transparent expressions that are either invariable or allow little variation (cf. Sag et al. (2002) for decomposable idioms and their syntactic diagnostics). Facing the task of translation, these idiom types merit subtly different approaches to both capture their meaning and relate them to the SL and TL lexicons.

Idioms additionally possess pragmatic levels of meaning external to any semantic (de)composition or formal linguistic properties (Fillmore et al., 1988; Croft and Cruse, 2004). Speakers know how to use certain expressions in the correct linguistic context and social situation. This pragmatic knowledge relies on, but is not viewed, traditionally, as contributing to the semantic interpretation of the original idiomatic expression itself. Idioms thus raise the question, should this "extra-linguistic" knowledge[4] be included in a computational lexicon?

[4]Knowledge not included in individual words, rules of grammar, or principles of compositional semantics

4.2 Idiom Translation Strategies

In translating an idiom, the question arises as to whether the translator ought to convey directly the presence of the idiom by maintaining the original words in literal translations of the SL, while perhaps also including a historical and cultural explanation. Alternatively, a translator could choose to simply translate from their understanding of the non-literal, implied meaning, without seeking a comparable idiom in the TL—that is, in effect removing any trace of the idiom.

Our translator found the work of Baker (1992) a helpful guide in approaching this translation task. Baker's four strategies have been employed in several idiom translation studies (cf. Strakšienė, 2009 for English-Lithuanian and 2010 for English-Russian; Shojaei (2012) for English-Turkish; Akbari (2013) for English-Persian). These are characterized in the tables below with MD idiom examples, each with a Romanized, tokenized transliteration, an English token-by-token gloss, and their corresponding English translation:

1. **Translation to an idiom of similar meaning and form**. This strategy is used when the SL and TL possess idioms of the same general meaning and lexical and syntactic content. This correspondence may occur either because idioms are compositional and transparent in both languages (or "literal" (Fernando, 1996)); or because the idioms are both non-compositional and opaque, yet conventionalized (Nunberg et al., 1994), in both languages.

MD Source Text	إلى كانت دارك جاج متشيرش على الناس بالحجر
Transliteration	ila kana-t dar-ek jaj ma-t-chayar-ch a'la al-nassa bi-l-ah'jar
Gloss	if were-it house-your glass not-you-throw-(NEG) at the-people with-the-stones
English Translation	People who live in glass houses shouldn't throw stones

Table 1: Example of idiom-to-idiom translation with preservation of form and meaning

2. **Translation to an idiom of similar meaning but dissimilar form**. This strategy is used when idioms in the SL and TL possess the same meaning but utilize different lexical items to convey that meaning. This correspondence occurs if the syntax available to create idiomatic meaning in one language is unavailable in the other. Likely, the idioms in SL and TL are "semi-idioms" (Fernando, 1996).

MD Source Text	المزوق من برى كيف حاليك من داخل
Transliteration	al-mzawaq min barra kif hali-k min dakhil
Gloss	the-colorful from outside how situation-your from inside
English Translation	Looks can be deceiving

Table 2: Example of idiom-to-idiom with shared meaning but different form

3. **Translation by paraphrase**. Paraphrase, the most common method to translate idioms (Shojaei, 2012), is employed when a match between SL and TL does not exist, or when it may be infelicitous to use idiomatic language in TL when it is felicitous in SL. This method is often used when idioms are opaque and non-compositional semantically; as such the more pragmatic meaning is translated. This occurs for "pure-idioms" (Fernando, 1996).

4. **Translation omitted**. Idioms are omitted from TL although present in SL for those cases when there are no equivalent expressions between the two languages, or when the SL meaning cannot be easily paraphrased for stylistic or pragmatic reasons.

Baker's four strategies work in a successive manner with the ultimate goal of preserving idiom meaning in the TL. Ideally, translators employ strategy (1), whereby the SL and TL possess idioms similar in both

MD Source Text	الا ربحتو ها وجهي
Transliteration	ila rbah't-u ha wajh-i
Gloss	if won-you here face-my
English Translation	There is no way you are going to win

Table 3: Example of idiom-to-[non-idiom] paraphrase

MD Source Text	گالو كحاز گالو ضهر الحمار قصير
Transliteration	gal-u kh'az gal-u dhar al-h'mar qssir
Gloss	told(he)-him move_over told(he)-him back the-donkey short
English Translation	Omission (1st guy: "make room", 2nd guy: "there's no room to spare")

Table 4: Example of idiom omitted, not translated

meaning and form. Given that this is not always possible, translators are instructed to proceed down the list of strategies to accomplish their task, as our translator did.[5]

4.3 MD Idioms: Preliminary Results

MD idioms, as found in the informal texts of social media and the commentary posts of our collection, includes intentionally playful wording, rhymes, and different types of figures of speech. The sentence structure of idioms may also be uncommon or unusual, though easy to understand. The idioms identified in our collection are generally non-transparent, i.e., their meaning cannot be derived from the words they contain. The wording of these idioms also generally does not allow for vocabulary substitutions or omission. Our translator observed that it is not uncommon for MD idioms to be structured as two-part, action-reaction expressions, indicating if X happens, then Y happens, resembling the two-part English construction, The X-er the Y-er.

With our work on creating a MD Lexicon, we isolated a total of 94 sentences with idiomatic expressions. Out of these, 52 were unique or existed in a single form. Of the remaining, two idioms occurred in nine different forms, one in six different forms, two in four different forms, and five in two different forms. After removing duplicates, we ended up with 62 unique idioms. A breakdown of which of the above translation types these idioms fall into is given below.

- Type 1 Translation to an idiom of similar meaning and form: No such cases in this specific collection were found. We have however found such examples in other data we collected (see Future Work).
- Type 2 Translation to an idiom of similar meaning but dissimilar form: In our example set, nine idioms out of the set of 62 unique idioms fell into this category.[6]
- Type 3 Translation by paraphrase: This category formed the overwhelming majority with a total of 51 out of 62 idioms.
- Type 4 Translation omitted: In our example sentences, we had two such cases.

The prevalence of translation by paraphrase is not surprising, given that this is the most common strategy cited in other translation efforts and given the rather extreme typological differences between English and MD. However, translation by paraphrase likely entails that potentially subtle nuances of meaning of

[5] Though Baker further identifies a fifth strategy for idiom translation, compensation, whereby idiomatic meaning is recreated elsewhere in the translated text, we do not make use of this strategy as our focus is on sentence-level translation and not discourse-level translation.

[6] Occasionally two idioms in a language may be used to convey the same meaning, providing translators with further issues to consider. In MD, for example, هزك الما وضربك الضو and مشيتي فيها are used interchangeably (same underlying meaning) and would both translate adequately as *You're fresh out of luck.*, even though individual native MD speakers may differ in their use of one over the other.

the idiom are lost in translation; additionally, for MD in particular, the playfulness of wording and sound expressed through syntactic form is lost. Although it may be difficult to pinpoint precisely which elements are lost in translation, there is value in indicating which strategy is used in translation for resources. A record of a lost idiom is at least an acknowledgment of the fact that there are likely undetermined meaning differences between the SL idiom and TL paraphrase that can potentially be recovered.

5 Constructions

Another set of challenges for translation comes in the form of constructions. Under a Construction Grammar approach, a *construction* is defined as any pairing of form and meaning, and therefore includes not only traditional entries in lexicons, such as morphemes and individual words, but also more complex forms, such as partially lexically-filled or fixed as well as fully general linguistic patterns (e.g., the previously mentioned The X-er the Y-er pattern) (Goldberg, 2003). The latter linguistic patterns are of special interest in this work because these meaning-carrying forms must be first recognized in order to then be correctly interpreted or translated. Such forms have not been included in many traditional generative lexicons.[7] Translations proceeding on a word-by-word basis would fail to recognize, for example, the semantics of correlation carried by The X-er the Y-er. This is problematic because, like idioms, the constructional semantics may require translation into a different construction or compositional phrasing in another language. Although idioms can be considered one type of construction, there are additional types of constructions carrying non-compositional meaning that are generally not included in a listing of idioms because they are fully general linguistic patterns that can be flexibly and productively filled with a variety of different lexical items. *Constructicons* (inventories of constructions) have been developed in some languages as a resource to facilitate recognition, interpretation and translation of constructions (Fillmore et al., 2012; Torrent et al., 2014; Bäckström et al., 2014). Here, we provide a preliminary exploration of constructions that should be considered for an MD constructicon.

A native speaker of MD and author of this paper examined instances of particular motion and degree-related constructions drawn from the Abstract Meaning Representation (AMR) corpus (Banarescu et al., 2013; Bonial et al., 2018). We selected AMR instances of constructions because the corpus offers a data-driven (as opposed to seeking out instances of a construction; these are constructions that arose in sentence-by-sentence AMR annotation) variety of annotated examples, which also facilitates a sense of the relative frequency of particular constructions in a larger corpus. After the native speaker of MD examined these English examples, examples of what were thought to be instantiations of the same construction in MD were selected. In addition, translation was attempted of some of the English examples in order to evaluate what constructions may or may not exist in both English and MD.[8] The findings of this analysis are given below, listed by construction type.

CAUSED-MOTION Construction (English) specifies that a causer argument directly causes a theme argument to move along a path designated by a directional phrase (Goldberg, 1995):

Subject.Agent Verb Object.Theme Oblique.Path
He blinked the snow off of her eyelashes.
They booed him off the stage.
She sneezed the foam off the cappuccino.

This construction is quite productive in English, licensing a variety of different verbs that are not typically associated with motion semantics, as shown above. Precisely what constraints exist upon the compatibility of a particular verb within this construction remains under debate. This idiosyncratic semi-productivity is precisely what makes adding this construction to the lexicon—as opposed to adding on motion senses

[7] Recently, expanding from within a generative framework, Dorr and Voss (2018) propose multi-layered verb structures for a computational lexicon to support spatial language understanding.

[8] Ideally, an MD constructicon would be developed through a careful analysis of the language through the lens of construction grammar, thereby avoiding the inevitable bias that stems from considering correspondents to English constructions. We hope to complete such analysis in the future as we develop the needed expertise in both construction grammar and MD. It would also be fruitful to explore a semi-automatic approach to detecting constructions, like that of Forsberg et al. (2014).

to what are typically non-motion verbs (like *blink*)—an effective strategy for providing more systematic coverage for the semantics of this construction in a computational lexicon. An examination of MD shows that a version of the Caused-Motion construction[9] exists in MD, as shown in Table 5.

We noticed, however, that the productivity and constraints on the Caused-Motion construction (e.g., determining which verbs can be licensed within it), at this stage in our analysis of MD, appear to be distinct from English. In particular, while English allows for a wide variety of verbs that lexicalize how the motion is caused (e.g., *boo, sneeze, blink*), we have observed that, for the examples our translator provided from MD, the motion semantics was expressed in two parts: through a distinct motion verb and through another word expressing the manner of causation, as shown in two example sentences in Table 6.

MD	قطع الصدفة من القميجة
Transliteration	qtaa' al-sadfa min al-qamija
Gloss	cut(he) the-button from the-shirt
English Source	He cut the button off the shirt

Table 5: Example of Caused-Motion construction in MD

MD	رمش حتى حيد الثلج من شفاره	اعطس حتى طارت الكشكوشة من الكابوتشينو
Translit.	ramach h'ta h'ayad al-talj min chfar-u	a'tass h'ta tara-t al-kachkouch min al-cappuccino
Gloss	blinked(he) till removed the-snow from eyelashes-his	sneezed(he) till flew_out the-foam from the-cappuccino
English	He blinked the snow off his eyelashes	He sneezed the foam off of the cappuccino

Table 6: English Caused-Motion examples expressed in MD through a motion verb & a distinct mention of cause/manner.

MD	دخلت الماشينة للاكار تتغوت	دخلت النحلة للبيت تتزنزن
Transliteration	dakhl-t al-machina l-al-laguar ta-t-ghawat	dakhl-t al-nahla li-l-biyt ta-t-zanzan
Gloss	entered-it the-train to-the-station (PRS)-it-scream	entered-it the-bee to-the-room (PRS)-it-buzz
English Source	The train whistled into the station	The bee buzzed into the room

Table 7: Sound Emission Verbs are realized separately from Motion Verbs in MD, in contrast to English, in which the Intransitive-Motion construction can license the motion semantics of what is typically a sound-emission verb.

INTRANSITIVE-MOTION Construction (English) carries motion and path semantics and licenses a variety of non-motion verbs, including sound emission verbs that are especially prevalent (Goldberg, 1995) :

Subject.Theme Verb Oblique.Path
The bee buzzed into the room.
The train whistled into the station.

However, MD—like many *verb-framing languages,* (Talmy, 1985; Talmy, 2000)—requires the sound emission verb be realized in surface form separately from the motion verb, as in Table 7.

COMPARATIVE construction (English) expresses the equality or non-equality of two values on a scale. For inequality, comparatives can be realized with a separate degree-word mention of what is more or

[9] Arguably, this could be an instantiation of a Resultative construction instead of a Caused-Motion construction. Further elicitation is needed.

less, or this can be realized in a comparative form of an adjective: [10]

Subject.Item1 Copula (Degree word) Adjective.Property AdverbialPhrase.Item2
This building is newer than the one we saw earlier.
The orange cat is less intelligent than its grey friend.

In MD, we observe that comparatives can be expressed either using the elative form of the adjective followed by من , similar to MSA, or, instead, by using the base form of the adjective followed by على . Examples for these cases are presented in Table 8.

MD	هد الطوموبيل اغلى من الشي لاخر	هد الطوموبيل غالية على الشي لاخر
Transliteration	had al-tomobile aghla min al-chi l-akhor	had al-tomobile ghalia a'la al-chi l-akhor
Gloss	this the-car more_expensive than the-thing the-other	this the-car expensive over the-thing the-other
English Source	This car is more expensive than the rest.	This car is more expensive than the rest.

Table 8: Examples of MD Comparative construction

The Superlative construction is realized very similarly to English, as shown in Table 9.

MD	هذ هو اكبر متحف فالبلاد	هذ هي اجود عمارة شفنا
Transliteration	hada huwa akbar moth'af fi-l-ablad	hadi hiya ajwad i'mara chaf-na
Gloss	this it biggest museum in-the-country	this it best building saw-we
English Source	This is the biggest museum in the country	This is the best building we have seen

Table 9: Examples of MD Superlative construction

THE X-ER THE Y-ER Construction (or `covariational conditional') has a unique form in which degree-words are nested into phrases with *the* occurring at the beginning (Goldberg, 2003):

The higher you fly, the harder you fall.
The more you practice, the less you will have to think about it.

The construction conveys the correlation between two variables changing along distinct scales. This construction is another excellent example warranting the development of constructicons, since it is the form alone that conveys this meaning, without any explicit lexical-semantic marking of correlation. In our analysis, we find that English instances of The X-er the Y-er can be translated into at least three distinct construction in MD. The three constructions are given in 5, where each is used as a possible translation for the same English sentence exemplifying The X-er the Y-er. While all plausible translations for The X-er the Y-er, these constructions certainly have distinctions in meaning and usage as well. Further analysis is required to determine those nuances and the extent to which The X-er the Y-er might be an adequate English translation of all three of these constructions. Thus, our analysis of MD translations of English constructions has led to three new potential candidates for a MD constructicon.

6 Related Work

6.1 Moroccan Darija

There are a variety of other recent efforts to create computational resources for MD, not to mention other Arabic dialects. For example, Al-Shargi et al. (2016) create a morphologically annotated corpus of MD, which they use to train an automatic morphological analyzer, and Darwish et al. (2018) leverage a diacritized MD Bible to build a diacritization system for MD. Another area receiving attention from

[10] FrameNet Constructicon: http://www.icsi.berkeley.edu/pubs/ai/framenetconstructicon11.pdf

English Source	The less we feed him the more defiant he gets
MD 1	شحال ما نقصنا ليه في الماكلة وهو تيزيد يحيح
Transliteration 1	chh'al ma nqass-na li-h fi al-makla w-huwa t-y-zid y-h'ayah'
Gloss 1	as_much_as that decreased-we to-him in the-food and-he (PRS)-he-increases he-defies
MD 2	ماحدنا نقصو ليه في الماكلة وهو تيزيد يحيح
Transliteration 2	ma-h'ad-na naqss-u li-h fi al-makla w-huwa t-y-zid y-h'ayah'
Gloss 2	as_long_as-we decreased-(PL) to-him in the-food and-he (PRS)-he-increases he-defies
MD 3	كلما نقصنا ليه في الماكلة كلما تيزيد يحيح
Transliteration 3	kul-ma nqass-na li-h fi al-makla kul-ma t-y-zid y-h'ayah'
Gloss 3	all/every-what/that decreased-we to-him in the-food all/every-what/that (PRS)-he-increases he-defies

Table 10: Examples of the-Xer, the-Yer MD constructions

NLP researchers is automatic dialect/language identification, especially for code-switched data, since MD speakers frequently mix MD with other languages such as MSA, French, and English. Samih and Maier (2016a) create a 223,000 token corpus from Moroccan internet discussion forums and blogs and annotate it for use in code-switching detection experiments (Samih and Maier, 2016b).[11] Voss et al. (2014) also build a code-switching detection system for MD, focusing on MD text written in Latin script.

In addition to the aforementioned digital resources, there are a few print dictionaries for Moroccan, including a recent verb dictionary (El Haloui and Bowman, 2011) as well as an older dictionary in Latin script edited by Harrell and Sobelman (1966). The Moroccan Darija Wordnet (MDW) project (Mrini and Bond, 2017) is endeavoring to create a WordNet from the Harrell and Sobelman dictionary and link it with Princeton's WordNet (Fellbaum, 1998). MDW will eventually be released as part of the Open Multilingual Wordnet project (Bond and Foster, 2013).

6.2 Idioms

In the last decade, researchers have turned to the task of automatically detecting idioms in English texts, developing statistical measures that go beyond the tradition of relying on manually identifying expressions in terms of their syntactic and semantic idiosyncracies; instead, researchers have analyzed a wide range of actual usage patterns in texts (Fazly et al., 2009), and most recently using word embeddings (Peng and Feldman, 2017) and leveraging existing sources of idioms, such as in Wiktionary (Muzny and Zettlemoyer, 2013). Related research in idiom identification in non-English languages is also being conducted by native speakers of those languages, e.g., Hindi (Priyanka and Sinha, 2014), Italian (Vietri, 2014), Russian (Aharodnik et al., 2018), and Japanese (Hashimoto and Kawahara, 2008).

6.3 Constructicons

Though constructicons are intended to be language-specific, they are based on the notion of continuity between the grammar and the lexicon in all languages (Fillmore, 2008). Resources for constructicons that we are aware of (Bäckström et al., 2014; Torrent et al., 2014; Forsberg et al., 2014; Ohara, 2016) take the following approach (similar to our own here): they begin by comparing English constructions to known constructions in another language of choice. Such comparison offers a baseline of constructions that may exist in the language, elucidating equivalent, approximate, and divergent constructions in the process; also, it addresses the possibility of linking constructicon resources in a dictionary-like manner. Backström et al. (2014) elaborate potential Swedish correspondents for Berkeley's English constructicon,[12] noting that full equivalence is difficult between constructions due to their use of content, form, and pragmatic usage combined. In contrast, Torrent et al. (2014) focus on a set of known constructions in Brazilian Portuguese (the *Para Infinitive* family), and through a close study of how to link theses constructions to

[11] Per personal communication with the authors, the corpus is not currently being redistributed due to intellectual property right concerns.

[12] https://www1.icsi.berkeley.edu/~kay/bcg/ConGram.html

English constructions as established by Goldberg (2003), they develop guidelines to help future annotators identify valid Brazilian Portuguese constructions. Though these constructions do not find full equivalence between languages, this type of analysis can provide useful diagnostics for identifying cross-linguistic commonalities and innovations alike.

7 Conclusion and Future Work

In this paper, we explore the challenges of building a computational lexicon for Moroccan Darija (MD). Starting from MD source text found online, we report on the construction of a parallel translation corpus of MD-English sentences and the computational lexicon derived in the process with bilingual MD-English entries.[13]

The construction analyses that we have begun, while preliminary, demonstrate that future work in linking constructicons across languages will be useful for translation by revealing places where one construction can be translated into another construction of a relatively similar form (e.g., for English and MD, the Comparative and Superlative), or a construction may be translated into one or more constructions of dissimilar form (e.g., The X-er the Y-er), and places where a construction in one language could only be translated periphrastically in another language (e.g., the English Caused-Motion and Intransitive Motion constructions translate into MD paraphrases). Instead of considering what English constructions may be present in MD, future work in expanding the MD constructicon will consider what unique constructions may exist in MD, which may not be expressed as constructions in other languages like English.

References

Katsiaryna Aharodnik, Anna Feldman, and Jing Peng. 2018. Designing a Russian Idiom-Annotated Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Monireh Akbari. 2013. Strategies for translating idioms. *Journal of Academic and Applied Studies (Special Issue on Applied Linguistics) Vol*, 3(8):32–41.

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 Language Resources and Evaluation Conference (LREC)*.

Linnéa Bäckström, Benjamin Lyngfelt, and Emma Sköldberg. 2014. Towards interlingual constructicography: On correspondence between constructicon resources for english and swedish. *Constructions and Frames*, 6(1):9–33.

Mona Baker. 1992. *In Other Words: A Coursebook on Translation*. Routledge, United Kingdom.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, page 178–186.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, page 1352–1362.

Claire Bonial, Bianca Badarau, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Tim O'Gorman, Martha Palmer, and Nathan Schneider. 2018. Abstract Meaning Representation of Constructions: The More We Include, the Better the Representation. In *Proceedings of the 2018 Language Resources and Evaluation Conference (LREC)*.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. The Multidialectal Parallel Corpus of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdurahim, Ossama Obeid, Salam Khalifa, Fadhi Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Kristin E. Brustad. 2002. *The Syntax of Spoken Arabic*. Georgetown University Press.

[13] We expect to release these resources to the research community, with further documentation in addition to this paper.

William Croft and D. Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge University Press.

Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, Younes Samih, and Mohammed Attia. 2018. Diacritization of Moroccan and Tunisian Arabic Dialects: A CRF Approach. In *Proceedings of the 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*.

Bonnie Dorr and Clare R. Voss. 2018. The Case for Systematically Derived Spatial Language Usage. In *Proceedings of the NAACL 2018 Workshop on Spatial Language Understanding (SpLU)*.

Abdennebi El Haloui and Steven L. Bowman. 2011. *Moroccan Arabic Verb Dictionary*. Artisanal Treasures.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, page 61–103.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Chitra Fernando. 1996. *Idioms and Idiomaticity*. Oxford University Press. Google-Books-ID: 5lViAAAAMAAJ.

Charles Fillmore, Paul Kay, and Mary Catherine O'connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, page 501–538.

Charles Fillmore, Russell Lee-Goldman, and Russell Rhodes. 2012. The Framenet Constructicon. *Sign-based construction grammar*, page 309–372.

Charles Fillmore. 2008. Border conflicts: Framenet meets construction grammar. In *Proceedings of the XIII EURALEX international congress*, volume 4968.

Markus Forsberg, Richard Johansson, Linnéa Bäckström, Lars Borin, Benjamin Lyngfelt, Joel Olofsson, and Julia Prentice. 2014. From construction candidates to constructicon entries: An experiment using semi-automatic methods for identifying constructions in corpora. *Constructions and Frames*, 6(1):114–135.

Adele Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Adele Goldberg. 2003. Constructions: a new theoretical approach to language. In *TRENDS in Cognitive Sciences*, volume 7(5).

Richard Slade Harrell and Harvey Sobelman, editors. 1966. *A Dictionary of Moroccan Arabic: Moroccan-English English-Moroccan*. Georgetown University Press.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an idiom coprus and its application to idiom identification based on wsd incorporating idiom-specific features. In *Proceedings of the Empirical Methods for Natural Language Processing Conference (EMNLP)*.

Khalil Mrini and Francis Bond. 2017. Building the Moroccan Darija Wordnet (MDW) using Bilingual Resources. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing (ICNLSSP)*.

Grace Muzny and Luke Zettlemoyer. 2013. Automatic Idiom Identification in Wiktionary. In *Proceedings of the Empirical Methods for Natural Language Processing Conference (EMNLP)*.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Kyoko Hirose Ohara. 2016. Toward constructicon building for japanese in japanese framenet. *Revista Veredas*, 17(1).

Jing Peng and Anna Feldman. 2017. Automatic Idiom Recognition with Word Embeddings. In *Proceedings of the Annual International Symposium on Information Management and Big Data*.

Priyanka and R.M.K. Sinha. 2014. A System for Identification of Idioms in Hindi. In *Seventh International Conference on Contemporary Computing (IC3)*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. "Multiword Expressions: A Pain in the Neck for NLP?". In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'02)*.

Younes Samih and Wolfgang Maier. 2016a. An Arabic-Moroccan Darija Code-Switched Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Younes Samih and Wolfgang Maier. 2016b. Detecting Code-Switching in Moroccan Arabic Social Media. In *SocialNLP Workshop at International Joint Conference on Artificial Intelligence (IJCAI)*.

Amir Shojaei. 2012. Translation of Idioms and Fixed Expressions: Strategies and Difficulties. *Theory and Practice in Language Studies*, 2(6), June.

Gary F. Simons and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World, Twenty-first edition*. Online version: http://www.ethnologue.com.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual Spoken Language Corpus Development for Communication Research. *International Journal of Computational Linguistics & Chinese Language Processing: Special Issue*, 12(3):303–324.

Leonard Talmy. 1985. Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description*, 3(99):36–149.

Leonard Talmy. 2000. *Toward a cognitive semantics*, volume 2. MIT press.

Tiago Timponi Torrent, Ludmila Meireles Lage, Thais Fernandes Sampaio, Tatiane da Silva Tavares, and Ely Edison da Silva Matos. 2014. Revisiting border conflicts between FrameNet and Construction Grammar: Annotation policies for the Brazilian Portuguese Constructicon. *Constructions and Frames*, 6(1):34–51.

Simonetta Vietri. 2014. The Lexicon-Grammar of Italian Idioms. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Clare R Voss, Stephen Tratz, Jamal Laoudi, and Douglas Briesch. 2014. Finding Romanized Arabic Dialect in Code-Mixed Tweets. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Mohammad Ali Yaghan. 2008. "Arabizi": A Contemporary Style of Arabic Slang. *Design Issues*, 24(2):39–52.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, page 49–59, Stroudsburg, PA, USA. Association for Computational Linguistics.