Building the Moroccan Darija Wordnet (MDW) using Bilingual Resources

Khalil Mrini[1], *Francis Bond*[2]

[1]Ecole Polytechnique Fédérale de Lausanne, Switzerland
[2]Nanyang Technological University, Singapore
khalil.mrini@epfl.ch, bond@ieee.org

Abstract

Moroccan Darija is one of the Arabic dialects, a continuum of under-resourced vernaculars. We develop a Moroccan Darija Wordnet (MDW) using a bilingual Moroccan-English dictionary, from which we collect nearly 13,000 definitions and over 15,000 lemmas. A Moroccan alphabet is set to make the MDW user-friendly. We link the Moroccan-English definitions to the Princeton WordNet using a method that found matches for about 77% of these, and estimated accuracy using confidence scores. Over 2,300 Moroccan synsets were verified as a first step of manual validation and are now included in the MDW, which is released as part of the Open Multilingual WordNet.

Index Terms: Moroccan Darija, WordNet, Arabic Dialect, Language Resource, Under-resourced Language

1. Introduction

Moroccan Darija is one of many variants of the Arabic language that can be defined as "*informal spoken dialects that are the media of communication for daily life*" [1]. Ethnologue lists it as *Arabic, Moroccan Spoken* (ISO 639-3 `ary`) [2]. In the 2014 census by Morocco's Higher Planning Commission [3], it is reported that Morocco has around 33.6 million inhabitants, and that 90.9% of them speak Moroccan Darija. It is therefore spoken throughout the country, albeit with small regional differences. It is called a dialect in Morocco, and as such has no standard orthography or official alphabet.

This paper is about the development of the Moroccan Darija Wordnet (MDW). It is released as part of the freely available Open Multilingual WordNet (OMW) [4, 5], and is the first dialect to be included in it. It is not however the first wordnet for a dialect, as a wordnet for the Iraqi dialect (IAWN) [6] was developed.

In this paper, we first describe the WordNet and previous work in Languages Resources for Arabic Dialects and in WordNet linking. Then, a dictionary of Moroccan Darija to English [7] is used to get the vocabulary for the MDW, for which an alphabet is set. Finally, the Moroccan vocabulary is linked to the English-language Princeton WordNet [8].

2. Related Work

2.1. Language Resources and Work on Arabic Dialects

Arabic dialects remain considered as under-resourced languages [9] and limited work has been done around them in Natural Language Processing.

Cavalli-Sforza et al. [6] develop an Iraqi Arabic WordNet (IAWN) and a method to link Arabic dialects to the PWN, as well as to the Arabic WordNet (AWN) [10]. The assumption used to make the IAWN is that a dialect is close enough to its base language so that their respective wordnets would have similar structures. This assumption is not used for the development of the MDW.

Habash and Rambow [11] present a morphological analyser for Modern Standard Arabic (MSA), and then adapt it for Levantine Arabic with linguistic data from Jordan. Chiang et al. [12] also exploit the similarities between MSA and Levantine Arabic to study the parsing of the latter through experiments in sentence and grammar transduction. Zbib et al. [9] crowdsource large English-Egyptian and English-Levantine parallel corpora to create a Machine Translation model.

Belgacem et al. [13] build a vocal corpus for nine dialects (Morocco, Algeria, Tunisia, Egypt, Lebanon, Syria, Iraq, the GCC countries, Yemen) and then propose a model for automatic recognition of Arabic dialects using Gaussian Mixture Models. Their results show how complex it is to distinguish Arabic dialects as they form a continuum.

2.2. Moroccan Darija Language Resources

Tachicart et al. [14] develop the Moroccan Dialect Electronic Dictionary (MDED). MDED is a MSA-Moroccan bilingual dictionary of 15,000 entries, obtained by translating an MSA dictionary to Moroccan Darija, and by translating a Moroccan Darija dictionary to MSA. The Moroccan Darija dictionary used was the French-Moroccan bilingual "*Dictionnaire Colin d'arabe dialectal marocain*" [15]. The alphabet used in MDED for Moroccan Darija is the Arabic one.

Samih and Maier [16] introduce an annotated Moroccan Darija code-switched corpus influenced by MSA. They demonstrate its possible uses by using it to detect code-switching at token and text level in Moroccan social media [17].

2.3. The Princeton WordNet

The WordNet is a lexical database created in the Cognitive Science Laboratory of Princeton University. It was first a database for the English language known as the Princeton WordNet [18, 19]. It regroups words by meaning in synsets, which are unordered sets of synonyms. Its latest release (WordNet 3.0) contains more than 150,000 unique words and around 120,000 synsets and is available for use via a web browser[1].

The WordNet divides synsets into four main parts of speech: nouns, verbs, adjectives and adverbs. However, it does not reference function words like prepositions and determiners. Therefore, these are excluded from the Moroccan Darija Wordnet. The noun synsets are also connected through relationships such as hyponymy, hypernymy, meronymy and homonymy. Likewise, verb synsets can have relationships such as hypernymy, troponymy and entailment. The advantage of the Word-

---

[1]Accessible on http://wordnetweb.princeton.edu/perl/webwn

Net is that correctly connecting a new language's synsets to the existing synsets keeps most of these relationships relevant.

2.4. The Open Multilingual WordNet

Wordnets were created for other languages and linked to the Princeton WordNet to form the Open Multilingual WordNet (OMW) [5]. It is akin to a programming-ready multilingual dictionary mainly used in applications of Machine Learning and Natural Language Processing, as it can be accessed through the Python-based NLTK package [20].

At the start of the OMW, it contained 26 wordnets. As of the time of writing, it lists 34 open wordnets merged. Linking the MDW to the OMW will therefore link it to 34 other wordnets and to 33 other languages, including Afro-Asiatic languages such as Arabic [10], Indo-European languages like French [21] and East Asian languages like Chinese [22, 23].

2.5. WordNet-linking approaches

Most of the freely available wordnets use the *expand* approach [24], which is mapping lemmas of the new language to the existing PWN English synsets. An English-Moroccan dictionary therefore enables the use of this approach for the MDW. Wordnets such as the Thai [25] and Indonesian [26] ones use it, although it is recognised as an imperfect method [27]. The approach in this paper is similar to the one presented in [28]. It uses two WordNet-linking attempts, and confidence scores to gauge the accuracy of the links they establish.

There exists also the *merge* approach for wordnet construction, which consists of building an independent monolingual wordnet and then mapping it to the PWN using bilingual resources. The EuroWordNet [29] is a multilingual wordnet project that uses this approach and has introduced an interlingual index (ILI) [30] to merge ontologies between different languages. Bond et al. [31] propose a collaborative ILI to expand it to other languages. The merge approach is used by wordnets such as the Urdu one [32], as well as along with the expand approach for the Russian one [33].

3. Creating a Moroccan Darija Wordnet

The OMW is based on the English-language Princeton WordNet, and it is the wordnet with the biggest number of synsets. Therefore, it makes sense to look for English definitions of Moroccan words and then link them to the Princeton WordNet. Moreover, even though there are French-Moroccan dictionaries, using them could cause inaccuracies as the French WordNet is sometimes deemed not as semantically reliable as the English one. Therefore an English-Moroccan bilingual dictionary would be most useful.

3.1. The Bilingual Dictionary

The dictionary used as basis for the MDW is the 1963 *A Dictionary of Moroccan Arabic: Moroccan-English*, edited by Richard S. Harrell. The entries were compiled by Thomas Fox and Mohammed Abu-Talib. Their goal was to collect the words that make up the core everyday vocabulary used by Moroccans. The dictionary does not cover dialect variations or terms deemed too technical. The entries were collected by interviewing educated Moroccans in three cities: Rabat, the capital; Casablanca, the most populated city; and Fez, the second-most populated city, according to the 2014 census. The choice of cities therefore ensures that the Moroccan Darija vocabulary

banefsež same as *bellefžuž*
banka pl. *-t* bank *mša xerrež le-flus men l-banka.* He went and got the money from the bank.
banyu pl. *-yat, -wat* bathtub
baqa pl. *-t* bouquet (flowers)
baqaʕ used in the expr. *baqaʕu li-llah* said interjectionally upon hearing of the death of s.o. (implying a "that's-the-way-it-goes" idea)
baqi a.p. of *bqa*
baqiya pl. *-t* rest, remainder
baṛ ibuṛ v.i. **1.** to be left over *had s-selʕa baṛet-lna.* We have this merchandise left over. **2.** to be or become an old maid *dak l-bent ġadi tbuṛ tul ʒayatha.* That girl is going to be an old maid all her life.
    ¶ *ʕateq bayṛa* pl. *ʕwateq bayṛin* spinster

Figure 1: Example of entries in Richard S. Harrell's *A Dictionary of Moroccan Arabic: Moroccan-English*

collected is typical Moroccan urban speech, understood in most parts of the country.

The entries are Moroccan lemmas followed by their English definitions or translations. Examples of entries can be seen in Figure 1. These could be either their direct translations, or a definition for words that are more specific. Another type of entry has one or more Moroccan lemmas followed by "*same as*" and one or more Moroccan lemmas, which have each an English definition entry somewhere else in the dictionary. Verbal nouns are denoted as "*[verbal noun] v.n. of [corresponding verb]*" in a separate entry, or as "*[verb] v.n. [corresponding verbal noun]*" in the entry of the corresponding verb. The plural forms, feminine forms, active and past principles are denoted likewise. Only the Moroccan-English definitions and "*[...] same as [...]*" entries are considered for simplification purposes, as other forms are inflected ones and cannot be linked to the WordNet.

In total, the dictionary contains 12,923 Moroccan-English definitions, or Moroccan synsets, with 14,409 lemmas, to which 720 are added from the "*same as*" entries, making a total of 15,129 lemmas.

3.2. Alphabet

There are existing Arabic transliterations such as the one proposed in [34] and the romanized Arabic transliteration engine proposed in [35]. However, the Moroccan Darija Wordnet is meant to be user-friendly and therefore it must be largely inspired by the unofficial alphabet that is used in daily life by Moroccans. Moreover, there are Arabic letters not used in Moroccan Darija, such as ذ, ث, and Moroccan sounds that cannot be denoted by the Arabic alphabet, such as *ḷ* (*l-ḷuṛ*), *ṛ* and *ẓ* (*zeṛbiya*).

The alphabet used by the Moroccan-English dictionary is based on the Latin alphabet and differentiates between short and long vowels. It denotes sounds not present in the Latin alphabet by Arabic letters, unlike the unofficial alphabet, or by using dots below letters for emphatic letters. Emphatic letters in the Arabic alphabet are ض (*ḍ*), ط (*ṭ*), ظ (*ḍ* or *ṭ*) and ص (*ṣ*). Their non-

Dictionary Alphabet	ġ	ح	ž	ʔ	ع
MDW Alphabet	8	7	j	2	3

Table 1: Differences between the dictionary's alphabet and the one used in the Moroccan Darija Wordnet (MDW)

emphatic equivalents are respectively د (*d*), ت (*t*), ذ (/ð/) and س (*s*). The dictionary's alphabet is based on phonology and has a one-letter-one-sound rule.

The alphabet used in this paper is as close as possible to the unofficial Moroccan alphabet to make it readable for the Moroccan public. The phonology rule was kept, and the alphabet includes emphatic letters and distinguishes between short and long vowels. In keeping with the writing that Moroccans use in daily communications, the Arabic letters were replaced by their respective numerals. The letters that were changed are referenced in Table 1. All other letters (*a, ă, b, d, ḍ, e, f, g, h, i, ĭ, k, l, ḷ, m, ṃ, n, o, q, r, ṛ, s, ṣ, š, t, ṭ, u, v, w, x, y, z, ẓ*) remain the same. The correspondance between both alphabets is reversible. In future work, we would like to explore how to facilitate looking up a lemma in the MDW, as errors may be frequent given that the orthography is not set and depends on pronunciation.

KtbDarija [36] is an example of a project that tries to define a specific alphabet for Moroccan Darija. It defines an alphabet in Arabic and an equivalent one in Latin letters with accents, but with no numbers. It follows Moroccan phonology as close as possible in both versions. To do so, it also associates one letter with one sound.

4. Linking to the WordNet

The Moroccan-English definitions are connected to the WordNet with a linking method consisting of two attempts. Then, we validate the results manually.

4.1. WordNet-linking Method

To connect Moroccan-English definitions to WordNet synsets, those definitions are first separated and given a Moroccan Darija Wordnet ID. Then there are two attempts, with the pseudocode of the first one provided in Algorithm 1.

In both attempts, everything that is between parentheses is discarded, as it could be noise or explanations that did not contain the bulk of the definition. With regards to the format of the dictionary, in which verb definitions start with "*to*", the part-of-speech tag is assumed to be *verb* when that occurs and therefore the definition would only be connected to synsets referring to verbs (lines 3 to 5 in Algorithm 1). Moreover, Part-of-Speech (PoS) tags in the beginning of the definition are taken into account to search for the correct PoS tag in the WordNet (lines 23 and 27 in Algorithm 1).

Each definition is split in one or more sub-definitions by commas or semicolons (line 6 in Algorithm 1). Each word in a sub-definition is filtered and thrown out if it belongs to the English "*stop words*" list of NLTK [20], with the exception of words such as "*up, down, out, in, on, off*" which accompany a verb and change its sense. A stop word is kept in the exceptional case where there is only one word in the sub-definition (lines 10 to 22 in Algorithm 1). The first attempt is stricter than the second. The second attempt was necessary to pick up matches with the WordNet that the first attempt could not establish.

4.1.1. First Attempt

In the first attempt, when searching for matching WordNet synsets, the lemma search uses underscore ("_") to connect the words in each filtered sub-definition (line 23 in Algorithm 1). For instance, the verb "*2amen b-*", with English definition "*to believe in*", is recognised as a verb because the definition starts with *to*. The WordNet is queried for "*believe_in*" with the part-of-speech tag being verb. The query yields one synset.

If searching with words connected with an underscore does not give results, the search takes each word in a sub-definition and matches it with synsets to form one set of synsets per word (line 27 in Algorithm 1). The final set of synsets for the sub-definition is the intersection of each word's set of synsets (line 29 in Algorithm 1). The sub-definition is considered only if its set is non-empty. Likewise, the final set of synsets for the definition is the intersection each of its sub-definitions' set of synsets (line 33 in Algorithm 1). If the definition's final set is non-empty, the synsets are associated to it. Otherwise, the definition cannot be connected to WordNet in the first attempt. The confidence score given to each match is 1.0 divided by the number of synsets.

As an example, the Moroccan noun "*2aṣel*" has the English definition "*origin, lineage*". We query the WordNet for "*origin*" and "*lineage*", and we get respectively 6 and 5 synsets. These sets of synsets have as intersection 1 synset, which becomes the synset corresponding to this Moroccan noun. The confidence score given is 1.0.

4.1.2. Second Attempt

While the first attempt only considers one possible set of WordNet synsets per Moroccan synset, the second attempt considers more than one. The second attempt repeats the same WordNet-matching method as the first attempt, starting by joining words with an underscore, and then splitting them if no results are obtained. However, it considers each sub-definition as an independent definition. Therefore, this attempt returns the final sets of synsets of individual sub-definitions rather than their intersection.

For example, if a definition has two sub-definitions, each with a non-empty set of synsets, that definition is split into two Moroccan synsets, respectively associated with the two sets of synsets. It is the case of the Moroccan noun "*amir*" with English definition "*emir, prince*". The WordNet queries for the words "*emir*" and "*prince*" give 1 distinct synset each, and therefore do not overlap. Therefore this definition's two sub-definitions "*emir*" and "*prince*" are considered as two separate definitions, each with 1 link to the WordNet.

The confidence score given to each match in the second attempt is 0.7 divided by the number of synsets. Therefore the maximum confidence score here is 0.7, hereby penalising the flexibility of this second run. In the above example, the two sub-definitions, which are linked to 1 synset each, both have 0.7 as confidence score.

4.2. Results and Validation

The matches resulting from the WordNet-linking method are given in Table 2.

The two attempts have resulted in a total of 12,224 Moroccan synsets connected to the WordNet. The 2,936 unconnected Moroccan synsets are mostly specific words embedded in the Moroccan culture and thus are not present in the WordNet. We have also left a list of 1,877 verbal nouns which are associated

Attempt	Synsets linked out of the total of 12,923	Links with 1 synset out of attempt's links	Links with 2 synsets out of attempt's links
First	59.6% (7,704)	33.0% (2,540)	17.6% (1,355)
Second	17.7% (2,283), split into 4,520 sub-definitions	22.0% (998)	18.6% (825)

Table 2: Results of the two attempts of the WordNet-linking method for the Moroccan-English dictionary

to verbs in the dictionary. In future work, they could be associated to the Moroccan Darija Wordnet through an additional relation.

The confidence scores enable us to quantify the accuracy of a WordNet link for manual validation. The validation here was conducted by one of the authors, who is a native speaker of Moroccan Darija. We first selected the 2,540 Moroccan synsets that are linked to the WordNet with confidence score 1.0. This confidence score means that they were each matched to exactly 1 synset.

Each link between a Moroccan synset and a WordNet synset is validated or rejected using the lemmas and definitions in the Moroccan-English dictionary and in the WordNet. During the validation, 8.7% (221) of the synsets were rejected. A main source of the errors is that there are adjectives and adverbs that are not tagged as such in the dictionary. Another error in linking are concepts which definitions contain light verbs, such as *get* and *make*, or started with *kind of*.

The 2,319 (91.3%) correctly linked Moroccan synsets are added to the version of the Moroccan Darija Wordnet that is now part of the OMW. They correspond to 2,571 Moroccan lemmas. In the future, we aim at completing the validation for all WordNet links to the MDW and add the verified data consequently.

5. Conclusions

We propose a Moroccan Darija Wordnet (MDW), with an alphabet close to the informal one popularly used in Morocco while complying with a one-letter-one-sound rule. The MDW is released as an extension to the Open Multilingual WordNet and is linked to the Princeton WordNet with a lexicon from a bilingual Moroccan-English dictionary. The dictionary has 12,923 Moroccan-English definitions, totalling 15,129 lemmas.

The dictionary is connected to the WordNet using two synset-connecting algorithms. Both attempts ignore English stop words with the exception of one-word definitions and separate verbs from nouns, adjectives and adverbs. The first one queries the WordNet for synsets and tries to find the common ones between words, cutting definitions into sub-definitions by commas or semicolons. The second one allows for more than one Moroccan synset per Moroccan-English definition. Both attempts attribute confidence scores to the matches they establish.

The first attempt has connected 59.6% of the the Moroccan definitions, and the second one has linked an additional 17.7%. They both resulted in 12,224 Moroccan-English definitions or sub-definitions connected to the WordNet. The 2,936 definitions remaining with no WordNet link are mostly words embedded in Moroccan culture and have no equivalent in the WordNet.

At the time of writing, all 2,540 Moroccan synsets with a 1.0 confidence score have been verified manually and 91.3% (2,319) passed the test. These are now included in the Moroccan Darija Wordnet. The latter will be enlarged as more synsets are validated.

Algorithm 1 First Attempt in WordNet Matching

1: **function** FIRST-MATCH($entry$) ▷ Where entry is an entry in the dictionary
2:     Let def be the definition in $entry$, id its ID, and pos its WordNet Part-of-Speech Tag
3:     **if** pos starts with "to" **then**
4:         $pos = wordnet.VERB$
5:     **end if**
6:     $subDefs = def.split(',',';')$
7:     Let $finalSynsetSet$ and $defSenses$ be empty arrays
8:     **for** $subDef$ in $subDefs$ **do**
9:         $words = subDef.split(' ')$
10:         **if** $pos == wordnet.VERB$ **then**
11:             **for** $word$ in $words$ **do**
12:                 **if** $size(subDef) > 1$ and $word$ in $stopwords$ except $(up, down, out, in, on, off)$ **then**
13:                     Remove $word$ from $words$
14:                 **end if**
15:             **end for**
16:         **else**
17:             **for** $word$ in $words$ **do**
18:                 **if** $size(subDef) > 1$ and $word$ in $stopwords$ **then**
19:                     Remove $word$ from $words$
20:                 **end if**
21:             **end for**
22:         **end if**
23:         $union = wordnet.synsets('\_'.join(words), pos = pos)$
24:         **if** $union$ is empty **then**
25:             Let $subDefSenses$ be an empty array
26:             **for** $word$ in $words$ **do**
27:                 Append $wordnet.synsets(word, pos = pos)$ to $subDefSenses$
28:             **end for**
29:             $union$ = intersection of non-empty sets in $subDefSenses$
30:         **end if**
31:         Append $union$ to $defSenses$
32:     **end for**
33:     $finalSynsetSet$ = intersection of non-empty sets in $defSenses$
34:     **return** $finalSynsetSet$
35: **end function**

6. References

[1] N. Y. Habash, "Introduction to arabic natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1–187, 2010.

[2] M. P. Lewis, G. F. Simons, and C. D. Fennig, Eds., *Ethnologue: Languages of the World*, 18th ed. SIL International, 2015, http://www.ethnologue.com/18/.

[3] H. C. P. Haut Commissariat au Plan du Maroc, "Recensement de la population," 2014.

[4] F. Bond and K. Paik, "A survey of wordnets and their licenses," in *Proceedings of the 6th Global WordNet Conference, Matsue, Japan*, 2012, pp. 64–71.

[5] F. Bond and R. Foster, "Linking and extending an open multilingual wordnet," in *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013. Sofia*, 2013, pp. 1352–1362.

[6] V. Cavalli-Sforza, H. Saddiki, K. Bouzoubaa, L. Abouenour, M. Maamouri, and E. Goshey, "Bootstrapping a wordnet for an arabic dialect from other wordnets and dictionary resources," in *Computer systems and applications (aiccsa), 2013 acs international conference on*. IEEE, 2013, pp. 1–8.

[7] R. S. Harrell, "A dictionary of moroccan arabic: Moroccan-english," *Georgetown University Press*, 1963.

[8] C. Fellbaum, "Wordnet: An electronic lexical database," 1998, mIT Press.

[9] R. Zbib, E. Malchiodi, J. Devlin, D. Stallard, S. Matsoukas, R. Schwartz, J. Makhoul, O. F. Zaidan, and C. Callison-Burch, "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*. Association for Computational Linguistics, 2012, pp. 49–59.

[10] W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, M. Bertran, and C. Fellbaum, "The arabic wordnet project," in *Proceedings of International Conference on Language Resources and Evaluation*, 2006.

[11] N. Habash and O. Rambow, "Magead: a morphological analyzer and generator for the arabic dialects," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 681–688.

[12] D. Chiang, M. Diab, N. Habash, O. Rambow, and S. Shareef, "Parsing arabic dialects," in *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[13] M. Belgacem, G. Antoniadis, and L. Besacier, "Automatic identification of arabic dialects," in *Proceedings of International Conference on Language Resources and Evaluation*, 2010.

[14] R. Tachicart, K. Bouzoubaa, and H. Jaafar, "Building a moroccan dialect electronic dictionary (mded)," in *Proceedings of the 5th International Conference on Arabic Language Processing*, 2014.

[15] Z. Iraqui-Sinaceur, *Le dictionnaire Colin d'arabe dialectal marocain*. Al Manahil, 1994.

[16] Y. Samih and W. Maier, "An arabic-moroccan darija code-switched corpus," in *In Proceedings of the International Conference on Language Resources and Evaluation*, 2016.

[17] ——, "Detecting code-switching in moroccan arabic social media," *Proceedings of SocialNLP@ IJCAI-2016, New York*, 2016.

[18] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[19] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[20] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly, 2009, (www.nltk.org/book).

[21] B. Sagot and D. Fišer, "Building a free french wordnet from multilingual resources," in *Proceedings of the Sixth International Language Resources and Evaluation (LRECâĂŹ08), Marrakech, Morocco*, 2008.

[22] S. Wang and F. Bond, "Building the chinese open wordnet (cow): Starting from core synsets," in *Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 10–18.

[23] C.-R. Huang, S.-K. Hsieh, J.-F. Hong, Y.-Z. Chen, I.-L. Su, Y.-X. Chen, and S.-W. Huang, "Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure," *Journal of Chinese Information Processing*, vol. 24, no. 2, pp. 14–23, 2010, (in Chinese).

[24] P. Vossen, "Building wordnets," 2005, accessed: 2017-08-07. [Online]. Available: http://www.globalwordnet.org/gwa/BuildingWordnets.ppt

[25] S. Thoongsup, K. Robkop, C. Mokarat, T. Sinthurahat, T. Charoenporn, V. Sornlertlamvanich, and H. Isahara, "Thai wordnet construction," in *Proceedings of the 7th workshop on Asian language resources*. Association for Computational Linguistics, 2009, pp. 139–144.

[26] D. D. Putra, A. Arfan, and R. Manurung, "Building an indonesian wordnet," in *Proceedings of the 2nd International MALINDO Workshop*, 2008, pp. 12–13.

[27] C. Fellbaum and P. Vossen, "Challenges for a multilingual wordnet," *Language Resources and Evaluation*, vol. 46, no. 2, pp. 313–326, 2012.

[28] K. Mrini and M. Benjamin, "Towards producing human-validated translation resources for the fula language through wordnet linking," in *Proceedings of the Workshop on Human-informed Translation and Interpreting Technology, held in conjunction with RANLP 2017*, Varna, Bulgaria, 2017, pp. 58–64.

[29] P. Vossen, "Introduction to eurowordnet," *Computers and the Humanities*, vol. 32, no. 2-3, pp. 73–89, 1998.

[30] C. Fellbaum and P. Vossen, "Challenges for a global wordnet," in *Online Proceedings of the First International Workshop on Global Interoperability for Language Resources*, 2008, pp. 75–82.

[31] F. Bond, P. Vossen, J. P. McCrae, and C. Fellbaum, "Cili: the collaborative interlingual index," in *Proceedings of the Global WordNet Conference*, vol. 2016, 2016.

[32] A. Zafar, A. Mahmood, F. Abdullah, S. Zahid, S. Hussain, and A. Mustafa, "Developing urdu wordnet using the merge approach," in *Proceedings of the Conference on Language and Technology*, 2012, pp. 55–59.

[33] V. Balkova, A. Sukhonogov, and S. Yablonsky, "Russian wordnet," in *Proceedings of the Second Global Wordnet Conference*, 2004.

[34] N. Habash, A. Soudi, and T. Buckwalter, "On arabic transliteration," in *Arabic computational morphology*. Springer, 2007, pp. 15–22.

[35] A. Chalabi and H. Gerges, "Romanized arabic transliteration," 2012.

[36] T. Daouda and N. Regragui, "Qyas kïbdarija : Projet pour un double standard pour l'ecriture de l'arabe marocain ou darija," 2012, (in French). [Online]. Available: http://www.ktbdarija.com/