An OpenNMT Model to Arabic Broken Plurals

Elsayed Issa
University of Arizona, 845 N Park Ave, Tucson, AZ 85719, USA
elsayedissa@email.arizona.edu

Abstract

The Arabic Language creates a dichotomy in its pluralization system; therefore, Arabic plurals are either sound or broken. The broken plurals create an interesting morphological phenomenon as they are inflected from their singulars following certain templates. Although broken plurals have triggered the interest of several scholars, this paper uses Neural Networks in the form of OpenNMT to detect and investigate the behavior of broken plurals. The findings show that the model is able to predict the Arabic templates with some limitations regarding the prediction of consonants. The model seems to get the basic shape of the plural, but it misses the lexical identity.

1 Introduction

The Arabic pluralization system creates an interesting phenomenon. The Arabic Language pluralizes its nouns and adjectives throughout morphologically linear as well as non-linear processes. While linear processes involve suffixation, the non-linear means involve infixation, that is, a change in the pattern of consonants and vowels inside the singular form. This phenomenon is distinguished by grammarians as broken plurals, and it is known for several Semitic languages including Arabic, Hebrew, and other Afroasiatic languages. Although several studies have examined Arabic broken plurals, this paper examines Arabic broken plurals using neural networks. The present paper attempts to build an OpenNMT neural network for training, testing and predicting broken plurals. It uses a large corpus of 2561 Arabic tokens. This attempt is twofold. It can help us approach this linguistic phenomenon using other methods, and it can explain or interpret the behavior of Arabic broken plurals templates. The importance of the present paper lies in detecting the behavior of not only broken plurals but also the behavior of sequences of consonant and vowels that make up these plurals. For instance, if the neural network can learn the singular pattern maʃʕal and the plural one mafaaʕil, but it predicts the words mænðạr (view) and manaaðịr (views) correctly while it fails to predict markaz (center) and maraakiz (center), which both have the same patterns, then other factors are to be examined to better understand the behavior of broken plurals. Additionally, this paper addresses the L2 acquisition benefits from the technology of neural networks in predicting the behaviors of L2 learners in their acquisition of Arabic broken plurals.

The paper is organized as follows. The introduction (section 1) introduces the research questions, describes the motivation behind the paper and establishes the argument. Section 2 lays out the concrete and necessary facts about the broken plurals and their patterns. Section 3 introduces the corpus of the study. Section 4 describes the methods used such as the OpenNMT as a general-purpose and attention-based seq2seq system. Section 5 reports the general performance of the experiment. Section 6 discusses and analyses the general performance of the experiment, presents the results, and discusses the impacts of new technologies – i.e. the OpenNMT – on second language acquisition. Finally, section 7, or the conclusion briefly summarizes the results.

Proceedings of the First International Workshop on Language Cognition and Computational Models, pages 22–30
Santa Fe, New Mexico, United States, August 20, 2018.
https://doi.org/10.18653/v1/P17

2 Arabic Broken Plurals

Two types of noun and adjective plural forms are present in the morphological system of Semitic languages. They are the sound (regular) plurals and the broken (irregular) plurals. Sound plurals, on the one hand, are formed by a linear process that involves adding the suffixes -uun/-iin in case of masculine nouns/adjectives, or -aat in case of feminine nouns/adjectives.

(1) Arabic Pluralization System

	Sing.	Pl.	Gloss
(a)	*muhandis*	*muhandisuun* (nom.)/-iin (acc./gen.)	(engineer)
	ṭaliba	*ṭalibaat*	(female student)
(b)	*qalb*	*quluub*	(heart)
	mænðạr	*manaaðịr*	(view)

In (1.a), the masculine singular noun *muhandis* (engineer) is pluralized as *muhandisuun* (engineers) in the nominative case or as *muhandisiin* (engineers) as in the accusative/genitive cases. The feminine singular noun *ṭaliba* (female student) is pluralized as *ṭalibaat* (female students). On the other hand, broken plurals are formed non-linearly by means of infixation or morphological transformations that involve internal consonant and vowel changes. In (1.b), the singular noun *qalb* (heart) is pluralized as *quluub* (hearts), a plural that involves a change in the pattern of the singular from *faʕl* (CVCC) to *fuʕuul* (CVCVVC). Similarly, the singular *mænðạr (view)* is pluralized as manaaðịr, and therefore, mapped on the pattern *mafaaʕil.* Ratcliffe (1990) concludes that there are 27 broken plurals patterns applicable to Modern Standard Arabic (MSA).

Therefore, Arabic broken plurals have stimulated the interest of several scholars. The non-linear treatment of template morphology of Semitic languages dates to McCarthy (1979, 1981, 1982 ...) and much more subsequent work. Hammond's (1988) contributes to the description of root-and-template morphology through the study of Arabic broken plurals. Moreover, in their in-depth paper, McCarthy and Prince (1990) have developed their theory of Prosodic Domain Circumscription where "rules sensitive to the morphological domain may be restricted to a prosodically characterized (sub-) domain in a word or stem." In the same vein, Ratcliffe's (1990) article aims at providing a framework for the analysis of Arabic morphology that involves the relationship between concatenative and non-concatenative morphology.

As far as the computation of broken plurals is concerned, Plunkett and Nakisa (1997) provide a connectionist model to the pluralization system of Arabic. They provide an analysis of the phonological similarity structure of the Arabic Plural system. In other words, they "examine whether the distribution of Arabic nouns is suited to supporting a distributional default in a neural network, by calculating a variety of similarity metrics that identify: (1) the clustering of different classes of Arabic plurals in phonological space; (2) the relative coherence of individual plural classes; and (3) the extent to which membership in a plural class can be predicted by the nearest neighbor in phonological space" (Plunkett and Nakisa, 1997). Their analyses show that the phonological form of the singular determines its sound plural. In their model, the distribution of Arabic singulars does not support a distributional default; however, their network performed well in (1) predicting plural class using the phonological form of the singular, (2) infecting singular to plural forms, (3) and generalizing the plural class prediction task to unseen words.

3 Corpus

The data consists of 2562 tokens extracted from a large contemporary corpus, provided with morphological patterns for both the singular forms and the plural forms. The data is organized into five columns as follows: lemma ID, singular form, singular pattern, broken plural form and broken plural patterns (Attia et al., 2011). The two columns of the singular form and the broken plural form are

extracted from the data, and then, they were prepared for the experiment using the R statistical language. The experiment is run for several times employing three different number of epochs; 10, 20 and 30 epochs.

4 Methods

Neural Machine Translation (NMT) has become a new evolving technology in the past few years. One of these NMTs is the OpenNMT (Open-Source Neural Machine Translation) which is a methodology for machine translation that has been "developed using pure sequence-to-sequence models" (Klein et al., 2017). This technology has become an effective approach in other NLP fields such as dialogue, parsing, and summarization. Also, Klein et al., (2017) maintain that OpenNMT was designed with three aims: (a) prioritize first training and test efficiency, (b) maintain model modularity and readability, (c) support significant research extensibility. In OpenNMT, four areas improve the effectiveness of the model. These four areas are gated RNNs such as LSTMs, large stacked RNNs, input feeding and test-time decoding (Klein et al., 2017). Although OpenNMT is built to handle sequence-to-sequence instances where it requires corpora of bilingual data to work, it can be used in other linguistic domains such as phonology and morphology.

As long as OpenNMT-py runs a neural machine translation that uses sequence-to-sequence long short-term memory (LSTM) to render a sequence of words into another sequence of words, this model uses OpenNMT as a tool that takes a sequence of broken plural letters and predicts them from a sequence of singular letters. The model deals with the non-linear morphological phenomenon of broken plurals as a machine translation problem where the input is the singular form, and the output is the plural form. The R code is used to vectorize singulars (as the input) and plurals (as the output); divide the data into one-third for validation, two-thirds for training, and 100 items for testing; and create log files for the three processes of OpenNMT; training, validation, and testing.

5 Results

Due to the small set of data, the model is run employing two stages. The first stage involves running the model for 10, 20 and 30 epochs without randomizing the data while the second stage covers the same number of epochs involving data randomization. The rationale behind this is training and testing the model for optimal results.

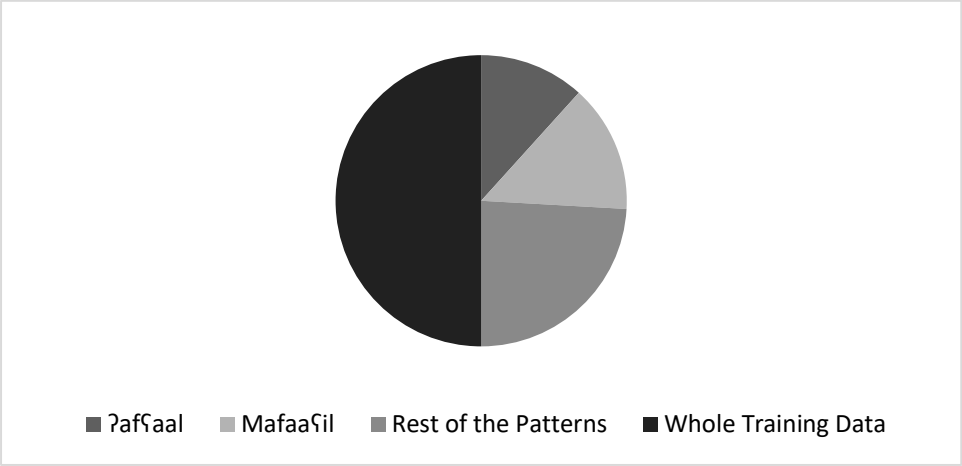
	Without Randomization			**With Randomization**		
	10 epochs	**20 epochs**	**30 epochs**	**10 epochs**	**20 epochs**	**30 epochs**
Prediction Average Score	-0.9190	-1.0311	-1.0610	-0.9733	-1.036	-1.0598
Validation Accuracy	55.9242	50.6183	51.4165	62.3264	62.3264	58.3398

Table (1) shows that the best results that characterize the performance of the model are at epochs 10 and 20 with a randomized data as well as 10 epochs without data randomization in case of the best prediction average score. The decline in the validation accuracy and the training accuracy can be due to: (1) the small amount of data in the corpus, (2) the small number of templates that the model learns. In addition, one assumption that validation accuracy is lower than training accuracy is the overfitting, meaning that the model learned particulars that help a better performance in the training data that cannot be applied in a large data. This, in fact, results in poor performance. Therefore, the model is run using a different number of epochs with randomization and without randomization to try to overcome the problem of overfitting.

6 Discussion and Future Work

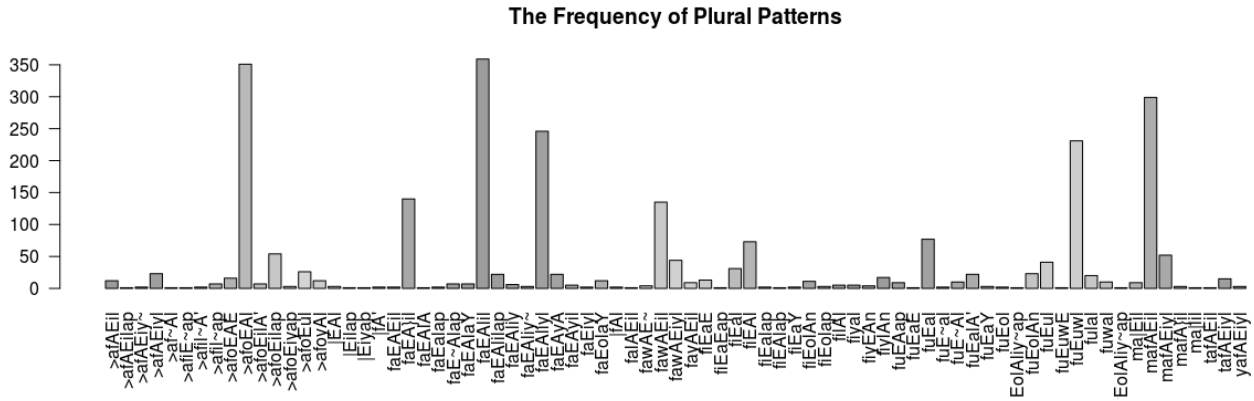
Based on these results, several points will be addressed. First, the examination of the training data shows that the data consist of 1641 observations which are divided into three categories. These involve the broken template *ʔafʕaal* with a frequency of 384 tokens, the template *Mafaaʕil* with a frequency of 466 tokens and 791 tokens for the rest of other templates. The frequency of the measures in the training data is shown in figure (1) below. It shows that the two patterns (*ʔafʕaal* and *Mafaaʕil*) constitute more than the half of the training data; therefore, the data predicted by the model will be greatly affected by these two patterns.

Figure 1. The Frequency of Patterns in the Training Data



Second, the examination of the database shows that the 2562 tokens are distributed among 124 singular templates and 77 plural templates. Figure (2) below shows the most occurring patterns among the plurals ones which repeat more than 100 times. Therefore, seven templates constitute 68.73% of the database, and subsequently, they have a great effect on both processes of training and prediction. Amongst these templates are the above mentioned two templates which are found to be dominant in the training data.

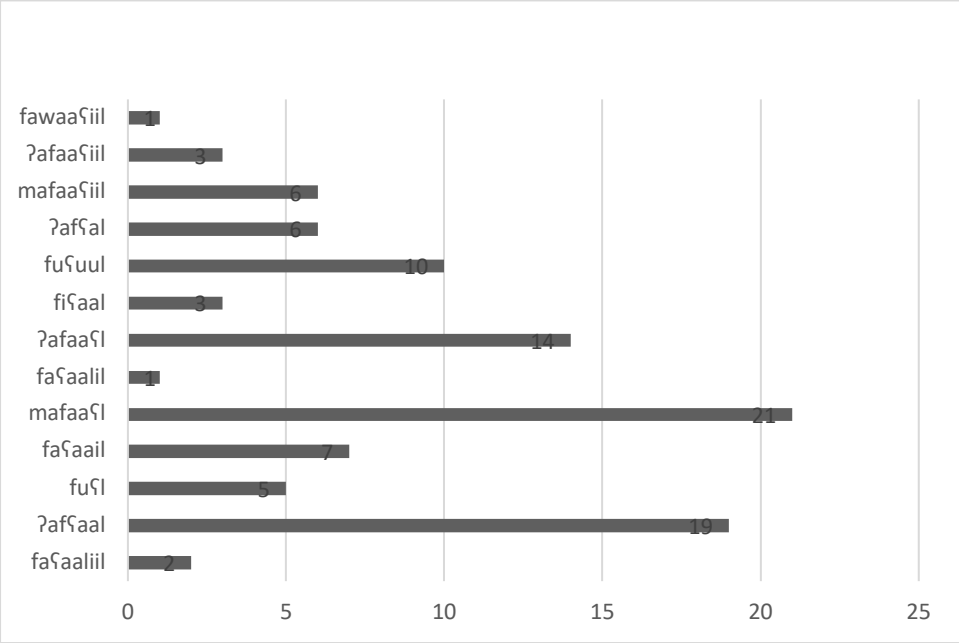
Figure 2. The Frequency of Arabic Broken Plurals in the Corpus of the Study



These seven frequent measures include >afoEAl (ʔafʕaal) which occurs 351 times, faEAil (Faʕaail) which occurs 140 times, faEAlil (Faʕaalil) which has 359 tokens, faEAliyl (Faʕaaliil) which occurs 259 times, fawAEil (Fawaaʕil) which has 299 tokens, fuEuwl (FuʕuwI) with 231 frequent tokens, and

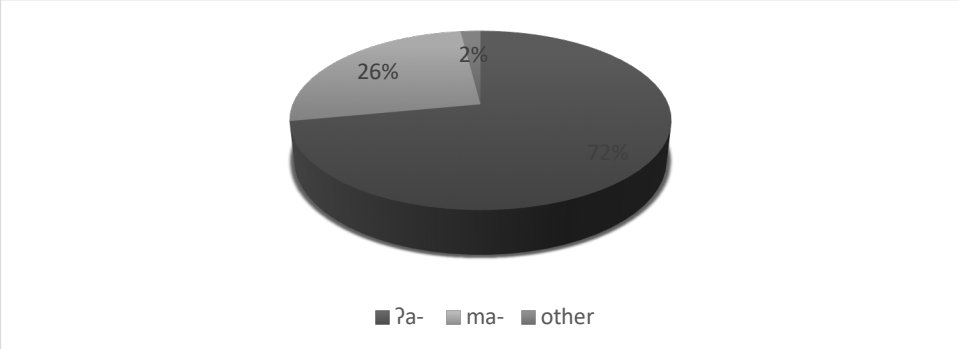
mafAEiyl (MafaaʕiiI) which occurs 299 times. The predicted data demonstrate 100 predicted plurals which consist of 13 different templates. The frequency of these predicted templates is shown in figure (3) below. The most salient templates are *ʔafʕaal* and *ʔafaaʕl* which start with the voiceless glottal stop /ʔ/ and *mafaaʕl* and *mafaaʕiil* which begin with the prefix /ma-/.

Figure 3. The Frequency of the Plural Templates in the Predicted Data



Considering this information in addition to the information introduced earlier about the frequency of *ʔafʕaal* and *MafaaʕiI* in the training data, it can be inferred that the predicted data is highly influenced by the two prefixes ʔa- and ma-. Accordingly, the data show that all the predicted plurals by the model involve templates that start with either the glottal stop /ʔ/ or the prefix ma-. The prediction of the data demonstrates an interesting phenomenon. Although the number of the template *MafaaʕiI* outnumbers the *ʔafʕaal* in the training data as illustrated in figure (1) above, the model predicts the templates with the prefix ʔa- more than the prefix ma-. This is, in fact, due to the frequency of patterns starting with the ʔa- prefix as illustrated in figure (3) above. Also, there are many templates, which do not belong to either ʔa- or ma- tokens, are assigned patterns starting with these two prefixes. Another interesting phenomenon is that the model can predict the structure of the pattern correctly in more than of the 60% of the predicted data. However, the model always changes one or two consonants and keeps the vowels in their slots within the template. In other words, the overall mapping of consonants and vowels to the patterns is successfully predicted as will be shown in the following discussion.

Figure 4. The Model Predictions with the Only ʔa- and ma- Prefixes.



The model succeeds in predicting most of the templates meaning that it manages to predict and map the vowels on the skeleton of the template while it fails to predict the consonants. The model's prediction of consonants ranges from predicting most of them, putting restrictions on the prediction of certain consonants such as gutturals and emphatics, and assigning divergent templates. The following discussion follows by examining the most salient patterns created by the model. The pattern *ʔafʕaal* is one of the most frequent plurals in the training as well as the predicted data.

(2) The Template ʔafʕaal

	Sing.	Pl.	Predicted Pl.	Gloss
(c)	Dawor /dawr/	>adowAr /ʔadwaar/	>arowAr /ʔarwaar/	(role)
	nawo' /nawʕ/	>anowAE /ʔanwaaʕ/	>anowAn /ʔanwaan/	(type)
	ku$ok /kušk/	>ako$Ak /ʔakšaak/	>awowAn /ʔawwaan/	(kiosk)
(d)	Sawot /ṣawt/	>aSowAt /ʔaṣwaat/	>awowAn /ʔawwaan/	(voice)
	DiEof /ḍiʕf/	>aDoEAf /ʔaḍʕaaf/	>arowAn /ʔarwaan/	(double)
	HawoD /ḥawḍ/	>aHowAD /ʔaḥwaaḍ/	>awowAn /ʔawwaan/	(basin)

According to the data shown in (2) above, the model is successful in predicting the plural pattern CVC.CVVC. However, it fails to keep the same consonants while it maintains the vowels. For instance, the first broken plural in (2.c) ʔa**d**waar (roles) is predicted as ʔa**r**waar where the voiced apical trill roll /r/ replaces the voiced apico-dental stop /d/. As for the second plural in (2.c), the model alternates the final guttural fricative /ʕ/ with the voiced alveolar nasal /n/. This can be attributed to the behavior of guttural in final position as there are other examples in the data that show the unpredictability of guttural sounds in final position. In (2.d), the model is successful in predicting the pattern; albeit with more changes in the consonants. It fails to predict the emphatics /ṣ/ and /ḍ/ whether in initial or final position. This may have two interpretations; the emphatics are either non-frequent in the distribution of Arabic consonants across the Arabic roots or their behavior restricts their predictability.

(3) The Template mafaaʕil

	Sing.	Pl.	Predicted Pl.	Gloss
(e)	maSonaE /maṣnaʕ/	maSAniE /maṣaaniʕ/	manA}iy /manaaʔii/	(factory)
	mafoSil /mafṣil/	mafASil /mafaaṣil/	manA}iy /manaaʔii/	(hinge)
	manoHaY /manḥii/	manAHiy /manaaḥii/	manA}iy /manaaʔii/	(prohibited)

Although the model is successful in predicting the template structure CV.CVV.CVC and the distribution of vowels within the template, it fails to predict the consonants. In (3.e), the model predicts the prefix ma- and the vowels where it could not predict the gutturals and the emphatics. Also, the model assigns the same predicted plural to three plurals. Also, the model inserts the /}/, the hamza /ʔ/ that has a seat in the middle of the word, into this template because it resembles another template which is *faʕaaʔil* as shown below.

(4) The Template faʕaaʔil

	Sing.	Pl.	Predicted Pl.	Gloss
(f)	Ea$iyrap /ʕašiiraa/	Ea$A}ir /ʕašaaʔir/	marA}iy /maraaʔii/	(tribe)
	ZaEiynap /ðaʿ ʕiinaa/	ZaEA}in /ðaʿ ʕaaʔin/	>awA}iy /ʔawaaʔii/	(wife)
	wadiyEap /wadiiʕaa/	wadA}iE /wadaaʔiʕ/	>awA}iy /ʔawaaʔii/	(deposit)

In (4.f), the model predicts the template as CV.CVV.CVC which fits two patterns; *mafaaʕil* and *faʕaaʔil*. It also predicts the seated hamza. According to the cases in (3 and 4) above, it seems that the

model learns the template, but it does not learn the distribution of the appropriate consonants on the template except for few consonants.

These observations can be attributed to examining the behavior of consonants. The frequency of certain consonants in the training data affects the model to predict specific consonants and rejects predicting the others. Therefore, there can be another experiment that examines consonants only. In other words, the experiment can involve only the consonantal tier of the broken plural. Since the Arabic morphology is interpreted in in terms of the CV-template, the study of the behavior, the frequency and the distribution of the consonants in the Arabic template can contribute to the prediction of the plural. According to the data used in this paper, it can be attested that certain consonants can occur more than other consonants in the template, i.e. the voiceless glottal stop /ʔ/. In the same manner, this proposes several questions about the distribution of some specific sounds, such as gutturals or emphatics, across the Arabic templates and the ability of neural networks, and hence, the human mind of predicting these sounds. If the predicted tokens are to be compared to plurals produced by children or L2 learners, the assumption of the difficulty of learning and predicting gutturals and emphatics will be attested. I assume that neural network is telling us about the difficulty of learning these sounds as human learners do.

(5) The Template fuʕal

	Sing.	Pl.	Predicted Pl.	Gloss
(g)	rasuwl /rasuul/	rusul /rusul/	>arowAr /ʔarwaar/	(prophet)
	tuhomap /tuḥmaa/	tuham /tuḥam/	>awAmim /ʔawaamim/	(accusation)

These are examples of how the model fails in predicting the template. Instead, it provides the template for the pattern *ʔafʕaal*. There are two assumptions for this prediction. First, the big frequency of the pattern *ʔafʕaal* contributes to this prediction. Second, the model maps the broken plural that has the pattern *fuʕal* to the singular pattern of the pattern *ʔafʕaal*, and therefore, it predicts the plural as *ʔafʕaal* as shown in the three examples in (5.g). For example, the broken plural *rusul* (prophets) in (5.g) can be analogized to the singular *dawr* (role) – this is the singular of the pattern *ʔafʕaal* – in (2.c) above. Hence, the model provides the predicted pattern *ʔafʕaal* (CVC.CVVC) to the broken plural with the pattern *fuʕal* (CV.CVC). All the predicted plurals for the plural with the template *fuʕal* have the template *ʔafʕaal* in the data predicted by the model. Therefore, the future work requires examining the broken plural in a larger corpus that also includes the Arabic sound plurals.

7 L2 Acquisition and Neural Networks

The conventional methods of teaching these broken plurals to L2 learners hold that there is a template for the plural to which the learner maps the stem of the singular into different syllable patterns by shifting the consonants of the singular form. Moreover, learners are told to use their "phonographic memory" to help them learn these patterns (Brustad et al., 2011, p. 30). For instance, given the singular form *dars* (lesson) and the plural template *fuʕuul*, they are asked to provide the plural form as follows:

(6) Mapping singular form to plural template:

fuʕuul (template)
duruu**s** (lessons)

They ignore the vowels and map the consonants to the root (**f-ʕ-l**) in the template; then they copy the vowels according to the melody that the template produces. The OpenNMT model is successful to some extent in capturing the melody of the template through assigning the vowels in their correct slots. However, the cases in which the model fails to capture the melody and assign the vowel, it predicts divergent plurals. Additionally, the model was successful in mapping the consonants and the vowels to the skeleton of the template as L2 learners can do and produce a correct template in approximately half

of the data. The failure of the model to predict gutturals and emphatics can be attributed to two factors. First, gutturals and emphatics might have less frequency than other sounds. Second, the model is behaving like an L2 learner who is learning according to the principle of the order of acquisition; namely, learning the easiest first, then the hardest. Probably, the model is addressing one of the arguments proposed by several studies that these sounds are the hardest to learn in the Arabic language. Therefore, more work should be done to address the benefits of neural networks technology in helping the acquisition of languages by foreign learners.

8 Conclusion

This paper attempts to look at Arabic broken plurals from the perspective of neural networks by implementing an OpenNMT experiment to predict the Arabic broken plurals. Broken plurals show an interesting phenomenon in Arabic morphology as they are formed by shifting the consonants of the syllables into different syllables patterns, which in turn, changes the pattern of the word. Therefore, they produce a melody besides changing the consonants. The paper seeks to describe these plurals using another method, i.e. OpenNMT, and detecting the way these patterns behave.

The findings show that several factors contributed to the predicted plurals. These include the frequencies of some templates as well as the distribution of consonants in the training data. Accordingly, the model predicts the templates most of the time with some alternations in the consonantal tier of the template, and it sometimes gets a different plural as a prediction of another plural. However, it succeeds to learn and predict the melodic tier of the template, i.e., it predicts the distribution of the vowels within the template. This prediction of vowels is similar to the way L2 learners learn to produce the broken plural given the singular form and the plural template. Therefore, another experiment can be implemented using the consonantal tier of the template for more inspection of these plurals.

Acknowledgements

I would like to thank Professor Michael Hammond for his help, excellent instruction, and insightful comments and suggestions.

References

Attia, M., Pecina, P., Tounsi, L., Toral, A., van Genabith, J. (2011). *Lexical Profiling for Arabic. Electronic Lexicography in the 21st Century*. Bled, Slovenia. Retrieved from: https://sourceforge.net/projects/broken-plurals/

Brustad, K., Al-Batal, M., Al-Tonsi, A. (2011). *Al-Kitaab fii Ta'allum al-'Arabiyya.* Part one. 3rd edition. USA: Georgetown University Press.

Hammond, M. (1988) Templatic Transfer in Arabic Broken Plurals. *Natural Language and Linguistic Theory*. (6) 247- 270.

Hammond, M. (2018). Neural Nets for Phonology and Morphology. R codes are retrieved from https://faculty.sbs.arizona.edu/hammond/ling696b-sp18/

Klein, G., Y. Kim, Deng, Y. Y., Senellart, J. & Rush, A.M. (2017) *OpenNMT: Open-source Toolkit for Neural Machine Translation.* ArXiv e-prints1701.02810.

McCarthy, J. (1982). "A Prosodic Account of the Arabic Broken Plurals." *Current Trends in African Linguistics*, 1.25. Retrieved from https://scholarworks.umass.edu/linguist_faculty_pubs/25

McCarthy, J. & Prince A. (1990) "Prosodic Morphology and Templatic Morphology." In Mushira Eid and John McCarthy (eds.) *Perspectives on Arabic Linguistics II: Papers from the Second Symposium on Arabic Linguistics*. Amsterdam: John Benjamins. 1-54.

Plunkett, K. & Nakisa, R., C. (1997) A Connectionist Model of the Arabic Plural System. *Language and Cognitive Processes*, 12:5-6, 807-836, DOI: 10.1080/016909697386691.