The efficiency of dropping vowels in Romanised Arabic script

Zeinab Kachakeche

Gregory Scontras

Richard Futrell
{zkachake, g.scontras, rfutrell} @uci.edu
Department of Language Science
3151 Social Science Plaza A
University of California-Irvine
Irvine, CA 92697-5100

Abstract

When Arabic speakers write in their dialect, they have the choice of using either the standard Arabic script or the non-standard Roman script. Arabizi writing is a new emerging writing system that Arabic speakers use to type their dialects utilizing Roman characters. Although Arabizi is not standardized, people have developed an efficient way to communicate through it. One phenomenon that emerged with this new system is vowel dropping. In this paper, we approach this phenomenon from the perspective of communicative efficiency. We study the informativity of short and long spellings of words and investigate whether the predictability of the word in certain contexts impacts whether the vowel is dropped in that word.

Introduction

Over the past decade, a lot of research has investigated the role of efficiency in human communication (Gibson et al., 2019; Jaeger & Tily, 2011; Pimentel et al., 2020; Piantadosi et al., 2012), among many others. While many have argued that human language is not efficient because of the types of ambiguity that any language holds, recent research has empirically tested the role of efficiency in human languages, proving it to play an important role in successful communication. We investigate the role of efficiency in a new and emerging writing system: Arabizi.

The rise of technological tools in the early 90s had a huge impact on language and communication. Not only did these tools facilitate and spread the English language across the globe, but they also didn't support languages that aren't written in Roman script. Arabic is one example of these languages. The lack of support for non-Roman scripts forced Arabic speakers who want to use such tools to create a new script for communication: Arabizi. Arabizi originates from the name "Arabic" and "Englizi" (the Arabic word for English), and it represents writing the Arabic language, mainly the various Arabic dialects, in Roman script. In addition to using Roman letters, Arabic speakers replace some phonemes that do not exist in English with numbers that are close in their orthography to the original Arabic letters. For example, the letter ع in Arabic is replaced with the number 3 and the letter ح is replaced with the number 7.

This new emerging system for writing Arabic is mainly used on social media platforms and SMS messaging and is far from standardization. The lack of standardization of such a system is due to a combination of factors. One huge factor is that Arabizi is the written form of the spoken dialect, and dialects differ widely across the Arab world. While many dialects are mutually intelligible, the varieties of pronunciation between them and within the same dialect affect the way people would write what they speak. Another factor that influences how people write their dialect is what other language they know that uses the Roman script. For example, an individual who knows French as a second language will write phonemes differently from another individual who knows English as a second language. For example, the sound /ʃ/ would be written in English as ⟨sh⟩ while in French as ⟨ch⟩. All of these factors lead to one interesting problem: there is not a one-to-one mapping between Arabic and Roman letters. This, however, provides us with unique grounds to study how efficiency affects the writing system at play.

In this paper, we investigate the efficiency of a particular property of the Arabizi writing system: vowel length. Following Mahowald et al. (2013), we test the hypothesis that the length of words is influenced by their predictability given the context it occurs in. We examine a set of 13 word pairs with long and short spellings, and we study whether the short spelling would occur in a more predictable context.

Background

According to Zipf (1936), the length of a certain word in language is determined by its frequency. Seventy-five years later, Piantadosi et al. (2011) challenged this idea by proposing that the average information content is a better predictor of word length than its frequency. The authors calculate word information in correlation to its linguistic context calculated through an n-gram model. They find that linguistic context is a better predictor of a word's length than its frequency. Building on the Piantadosi et al. (2011) results, Mahowald et al. (2013) considered pairs of long and short forms of English words and tested whether more predictable contexts affect the use of either form. They find that in more predictable contexts, shorter forms are more likely to be used. The authors support their corpus results by further conducting a behavioral study and obtaining the same effect. Their proposal is that the higher the information content of a word, the higher the surprisal rate of that word, and the less predictable that word is in a given

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

context. In the current paper, we follow the Mahowald et al. (2013) methodology in calculating the surprisal of short and long spellings of words as a measure of predictability, and we test whether it impacts the length of the words chosen.

In Arabic, there exist three vowels, /a/, /i/, and /u/, which may be long or short; the difference between the two forms of vowels is phonemic (Gordon, 2011). When writing these vowels in Arabizi, there is a lot of variation in how speakers represent the vowels. For example, the letter 'Yaa' in Arabic (ي), which is the long /ii/ vowel, could be written as ⟨ee⟩, ⟨ii⟩, ⟨ea⟩, ⟨y⟩ or shortened to ⟨i⟩, or ⟨e⟩, among other representations. Also, the short /i/ vowel, al-kasra, can be written as ⟨e⟩, ⟨i⟩, or dropped all together. This applies to all the pairs of short and long vowels in Arabic. The phenomenon of vowel dropping in Arabizi leads to high variability in writing words. One example is the word *kataba* 'to write' in Standard Arabic, which has three short /a/ vowels. *Kataba* can be written in dialectal Arabizi as ⟨katab⟩, ⟨ktb⟩, ⟨katb⟩, or ⟨ktab⟩. On the other hand, the word *kitaab* 'a book', which has a long /aa/ vowel in the second syllable, can be written in Arabizi as ⟨ktab⟩ or ⟨kteb⟩ depending on the dialect. In this paper, we are particularly interested in investigating the roots of this phenomenon, and whether efficiency plays a role in dropping vowels in Arabizi.

Many have studied the Arabizi writing system in an attempt to build Machine Learning models to recognize it (Tobaili, 2016; Shazal et al., 2020; Baert et al., 2020), or to create automatic translation systems between Arabizi and standard Arabic (Darwish, 2013; Bies et al., 2014) or between Arabizi and English (May et al., 2014). These studies mainly focused on building models that optimize the performance of machine learning systems since Arabizi remains under-studied in the field of Natural Language Processing. Approaching Arabizi from a socio-linguistic perspective, some have studied the linguistic features carried with this new emerging system. Some of these studies investigated vowel dropping in Arabizi writing, which is the topic of our current paper. Gordon (2011), for example, explores the orthography of Arabizi in the written Levantine dialect, and reports that short vowels might be dropped. Akbar (2019) investigates vowel and consonant deletion in Kuwaiti Arabizi using data collected from group e-conversations as well as private WhatsApp messages of 35 students, and finds that consonants are rarely deleted while short vowels are deleted more often. The authors also report that long vowels are slightly shortened. An example from Kuwaiti Arabic is illustrated in the paper that the spelling ⟨7aalıch⟩, which means 'your situation', has two vowels: a long /aa/ in the first syllable, and a short /i/ in the second syllable. Speakers showed to delete the short vowel in the second syllable and shorten the long vowel in the first syllable resulting in the word being written as ⟨7alch⟩. Sullivan (2017) explores orthographic variation in Lebanese Arabizi on Twitter, and reports the different writings of each of the Arabic alphabets in the data. In addition, the author reports that the most variable writing was that of the short vowels. Many of the mentioned papers provide explanations for the phenomenon of dropping vowels. One explanation is that Arabic speakers are sticking to the Arabic orthography when writing in Roman script. That is, when writing standard Arabic, speakers tend to drop the short vowels (which are represented as diacritics) when unnecessary. However, this doesn't explain the instances when speakers write the short vowels.

None of the papers aforementioned study the phenomenon of vowel dropping from an information theoretic perspective, nor explain this phenomenon from the lens of communicative efficiency. In this paper, we primarily focus on explaining this phenomenon using an information theoretic approach, following the Mahowald et al. (2013) study. We explore the role of context in determining word length, and consequently leading to vowel dropping when the vowels are not essential for successful communication. The variability in Arabizi writing could lead to lexical ambiguity. However, lexical ambiguity is easy to resolve given context, while writing the necessary vowels is more costly on the speaker's behalf because it takes more time to type. Consequently, it could be more efficient for the speaker to drop the vowels in writing because they know the listener will disambiguate. Moreover, if the context prepares the listener to perceive certain words (hence that word is more predictable in that context), then it is easier and more efficient to drop the vowels than not. In other words, if the vowel can be easily disambiguated by context, dropping the vowel will reduce cost on the speaker's behalf, and hence facilitate faster communication. However, speakers can only do this if the reduced form of the word is predictable enough, and can be disambiguated given the context.

Research questions and hypothesis

In this paper we investigate whether predictability given context affects the length of certain words by calculating the surprisal of short and long spellings of words as a measure of words' informativity. The surprisal of the words in given contexts can also be viewed as a measure of predictability: the higher the surprisal, the less predictable the word is. We expect to see that speakers use the shorter spellings of the words when the word is more predictable (and hence has low surprisal rate). We also test the impact of frequency on the length of the words. That is, if a word is more frequent, are Arabic speakers more likely to drop the vowels when writing Arabizi?

Methodology

Choosing short and long forms of the words

We follow the Mahowald et al. (2013) methodology in comparing the use of short and long pairs of Arabizi words with respect to predictability of the word. We used the Egyptian Arabic text chat conversations from the Arabic Treebank of the BOLT (Broad Operational Language Translation) project through the Linguistic Data Consortium (LDC) to extract the most frequent words used by Egyptian speakers in the data set. The data set consists of 157,569 lines and 694,910 tokens,

short variant	long variant	translation
ana	anaa	I, me
msh	mesh	not, will not
bs	bas	just, only, stop
kda	keda	like this
mn	min	from
tyb	tayeb	okay
3shan	3ashan	because
y3ni	ya3ni	meaning that, that means
knt	kont	I was
m3	ma3	with
hwa	howa	he
tmam	tamam	perfect
kman	kaman	also

Table 1: This table shows the Arabizi pairs used in training the model. As mentioned above, many of these pairs were function words such as pronouns and prepositions.

of which 120,129 are unique. The data is Egyptian Arabic written in both Arabic script and Arabizi, and is not annotated. This Arabic data set was collected from SMS and chat messages. We first pre-process the data to get it to a usable format. We then classify the data into Arabic script and Arabizi, and we only consider the Arabizi text for our analysis. Our classification technique is simple, where we only consider the text written in Roman characters as Arabizi using a Unicode function in Python (`encode().isalnum()` function). We were able to do that since we already know, based on the LDC website, that Arabizi is the only Romanised script in the data. The new Arabizi data set consists of 103,141 lines and 460,079 tokens, of which 81,857 are unique. We then generate a list of the most frequent words in the Arabizi data set. Based on the most frequent words, we manually choose 13 pairs of words with short and long spellings. For example, the word /ana:/ which is the equivalent to pronoun 'I' or 'me' in English, is written as [anaa] or might be alternatively written as [ana] in its reduced form after the long vowel is dropped. The list of chosen pairs of words are shown in table 1. Because of the nature of the data set, some of the pairs of words are function words (ex. pronouns and prepositions).[1] We are aware that there might be multiple ways to write some of these words (For example, *anaa* could also be written as ⟨ane⟩, ⟨ani⟩, etc.) depending on the dialect. However, for the current paper, we only consider two spellings of each word for simplicity. For choosing the longer spelling of some of these words, we took advantage of the MADAR lexicon (Bouamor et al., 2018). The MADAR lexicon provides different variants of the words as they are pronounced in different Arabic-speaking cities. That is, the words in the lexicon are the most informative based on the pronunciation of each dialect. We used the pronunciations of words from Cairo city (to match the LDC Egyptian Arabic data set) to determine the longer spellings of words in the Egyptian dialect. For the words of which spellings from the MADAR lexicon didn't exist in the BOLT data set, we modified the spelling to the closest one possible that existed in the data set.

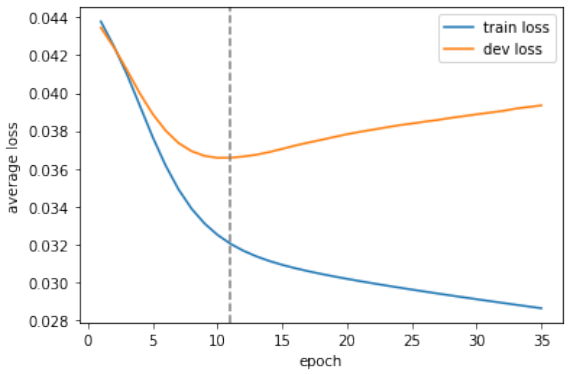

Figure 1: This plot shows the cutoff we used to determine how many epochs to keep training the model. On the x-axis is the number of epochs and on the y-axis is the negative log likelihood loss divided by the size of the data set (i.e., the number of bi-grams). Notice that around 11 epochs, the loss on the dev (development) set starts increasing, which is an indication of over fitting. The y-axis is the average loss, so even though the dev set is significantly smaller than the train set (10% and 90% of the data respectively), they are on the same scale.

Training an n-gram model

After choosing the appropriate list of pairs, we trained a neural n-gram model on the BOLT data set. We first split the data into train and development sets which are respectively 90% and 10% of the data. We trained the n-gram model using Pytorch. After many trials with different models, learning

[1]The list of pairs will be revisited in future extensions of this work

rates, parameters, and optimizers, the best model that achieved the smallest loss on the development (dev) set was a bi-gram model trained on 11 epochs and optimized with Adam (an algorithm for gradient-based optimization; Kingma & Ba, 2015) and an initial learning rate of 1e-5. During each epoch, the model trains on the entire train data set, and attempts to minimize the loss. We evaluate the improvements of the model by calculating the loss on the dev set. The loss function we used to measure the performance of the model is negative log likelihood. Optimally, we want the dev set loss to be as close to zero as possible in order to achieve the highest performance of the model. One common issue with training models is that the model achieves a small loss on the train data set, but doesn't do the same on the dev set. When this happens, it is an indication that the model doesn't generalize well to unseen data and this issue is commonly known as over fitting. To prevent our model from over fitting, we monitor the dev set loss at the end of each epoch. In our experimentation, we found that the model started over fitting after the 11th epoch, in which case we decided that 11 is the optimal number of epochs (see Figure 1). Training a tri-gram model took more time to learn (16 epochs), but didn't reach as low of a loss as the bi-gram model. Our model had a total of 18,758,181 trainable parameters (see Figure 2). We trained the model on a GPU for faster running time. Each epoch took an average of 38 seconds. We didn't use any smoothing techniques for our model because the model is implicitly smoothed. That is, the neural network model is unlikely to produce zero probabilities, but instead will produce extremely small probabilities.

Modules	Parameters
embedding.weight	8185700
linear1.weight	12800
linear1.bias	128
linear2.weight	10477696
linear2.bias	81857

Total Trainable Params: 18758181

Figure 2: A table showing the different parameters used in the model. Amongst the parameters, 18,758,181 were trainable. Some of the important hyper-parameters we fed to the model which determined the number of trainable parameters are embedding dimension which we chose to be 100, and the linear dimension which we decided to be 128.

Obtaining informativity measures

Using the bi-gram model, we calculated the average surprisal of each word given the context. Assuming that surprisal is a measure of informativity, the higher the surprisal rate of a certain word written in either spelling, the more information that word carries. Consequently, words with higher information measures are less predictable in the given context. Our hypothesis is that speakers are more likely to use shorter spelling variants of words in more predictable contexts. To test this hypothesis, we measure the surprisal (hence predictability) of the meaning of the word by combining the probability of that word occurring in either spelling in a given context. We calculate the surprisal using the following equation:

$$-1/N\sum_{i=1}^{N}\log P(W=w|C=c_i) \quad (1)$$

where N is the total frequency of some word w, and c_i are all possible context words. Since the bi-gram model we trained proved to be the most reliable, the context c_i is considered to be the one word preceding the target word w. The bi-gram model provides probabilities for all the possible utterances after every word. We add the probabilities of the independent spellings of each word, and take their log. We sum the log probabilities of the words to bring the probabilities to a more manageable scale, since computing just the probabilities will return very small numbers. We then average the log probabilities over the contexts in which the words actually occurred in the data set. We labeled the context as long if the long spelling occurred in that context, and we labeled it as short if the short spelling occurred in that context. We then took the average surprisal, represented by the log-probabilities, over the long and the short spellings of the words. We subtracted the average surprisal of the short spellings from the average surprisal of the long spellings for every pair of words. We plotted the results against the combined frequency of the short and long spellings of the words after taking their log (Figure 3). The results show that most of the average surprisal differences of the words lie above the y=0 line, which means that most of the time when writing Arabizi, speakers tend to use the shorter spellings more than the longer ones. If significant, these result could show that the average predictability of a meaning across certain contexts impact which spelling of the word was used.

We ran a mixed-effects logistic regression predicting word form (short vs. long) by surprisal of the word meaning with random intercepts for word meaning. The average surprisal for the long spellings of the words was 4.92, which is higher than that of the short spellings (4.51; $\beta = -0.111, z = -2.76, p < 0.01$). A t-test aggregating all the short spellings and long spellings of all of the words and comparing their average surprisal rates was also significant. ($t = 21.262, p < 0.001$). This suggests that the longer spellings of the words have higher information content than their shorter counterparts. We also performed a series of t-tests, one for each word spelling, to see if the mean surprisal difference of the short spellings and the long spellings is significant at the level of individual word forms. None of these differences were significant, likely owing to the fact that we had substantially less data for the individual word form analyses. Figure 3 also shows that there isn't a role for frequency in the choice of which word spellings to type ($r^2 = 0.29$, Spearman's $\rho = -0.483$, p= 0.097). In other words, the results from the figure suggest that the frequency of

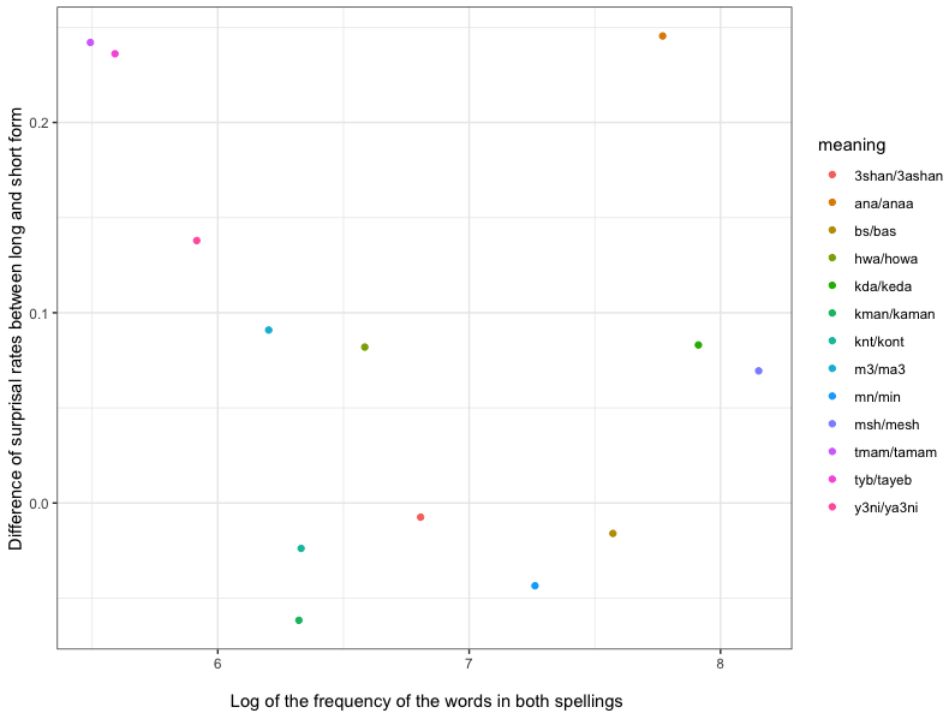

Figure 3: Difference in average surprisal between the long and short spellings of the words plotted against the log frequency of the word pairs combined.

the word might have a role in determining whether the vowel gets dropped or shortened, but information content (measured in surprisal) plays a larger role in determining word length.

Running the model on Twitter data

After obtaining our informativity measure, we ran the model on another set of data from Twitter that was unseen to the model. We extracted the Twitter data using API search. We pulled tweets that contained the 13 pairs of words we obtained the informativity scores for. To be consistent with the LDC data set, the tweets we pulled were based on the geographic location from Cairo city in Egypt. The data was originally ~ 6400 tweets. We filtered out the bi-grams where the first word was new and didn't exist in the training data set, because the model has not trained on these tokens and doesn't have the probabilities of the next words occurring after them. The remaining bi-grams were 4719. We then computed the difference in surprisal rates between the long and the short forms the same way we did above. The results we obtained are shown in Figure 4. We didn't plot the surprisal against frequency because we already assume the frequency of these pairs of words is similar according to the query we built to pull the tweets. By conducting a t-test for whether the word form (short vs. long) is affected by the surprisal of the meaning given context, aggregating over words, we find the expected effect ($t = 11.801, p < 0.001$). However, the effect is not significant in a mixed-effects logistic regression with random intercepts for word meaning($\beta = -0.03628, z = -0.586, p = 0.558$).

Discussion and future work

Our results from the first set of data show that word predictability is correlated with word length: The more the word is anticipated in a given context, the more likely the short spelling of a certain word is used (and hence the more likely speakers are to drop the vowels when typing Arabizi). Our study considers one type of context: linguistic context represented by one word immediately preceding the target word using our trained bi-gram model. These results suggest that context plays an important role in language production. Longer spellings are more costly, so speakers are more likely to use the shorter spellings whenever they can, unless the longer utterances carry necessary information. Hence, if the context provides enough information to disambiguate shorter utterances, speakers will give away less informative vowels for the sake of efficiency. What makes these results even more interesting is that neither the short spellings nor the long spellings in Arabizi are standard, and writing words may vary significantly from one person to another and from one dialect to another.

One distinction between the two data sets we used to train and test our model is that the training data and part of the test data comes from chat conversations and SMS messages while another part of the test data comes from public tweets. This distinction, while subtle, is very important; speakers privately chatting with each other might tend to drop vowels more because they have more background knowledge about their listener. On the other hand, people posting tweets publicly will tend to be more accurate in their language (and thus drop their

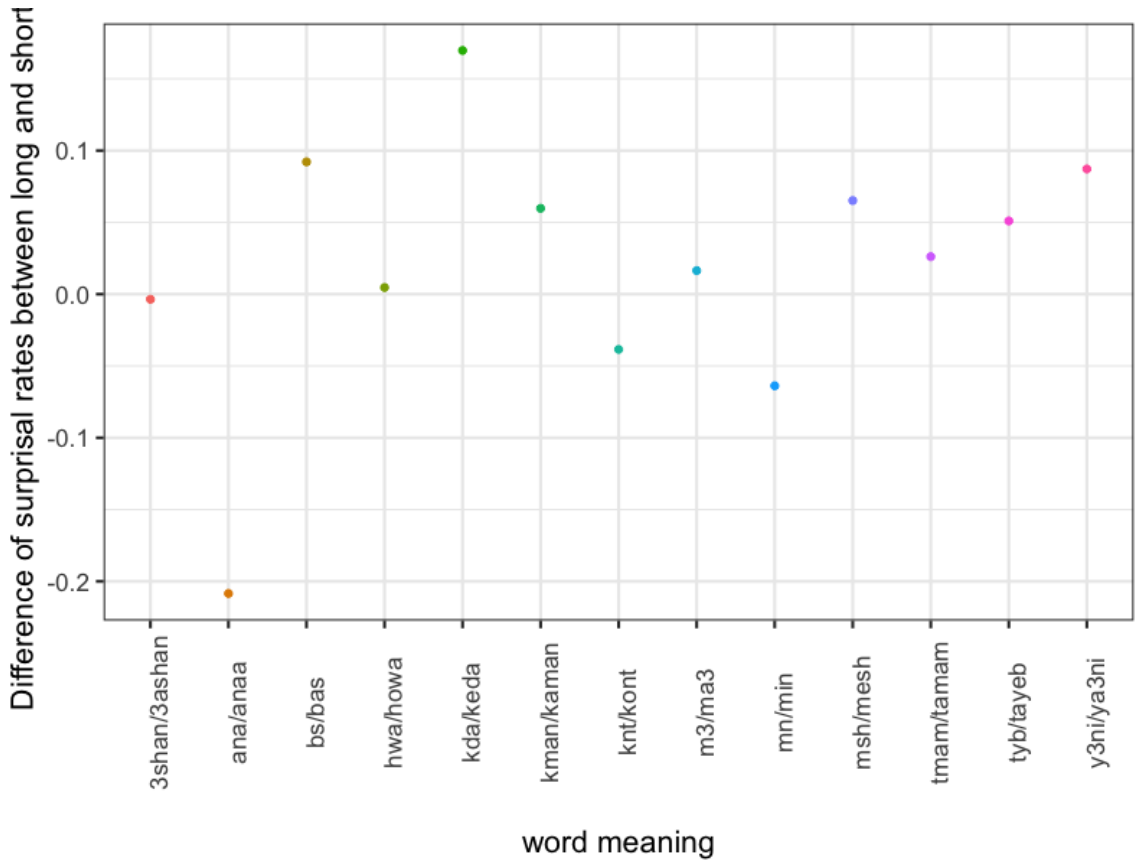

Figure 4: This plot shows the difference in surprisal between the long and short forms of words in Twitter data. The surprisal of the words is plotted against the word pairs rather than against frequency since we assume the word pairs to have similar frequencies in this data. With the Twitter data, we notice that the word pairs are used differently than in the private SMS chats in the training data. For example, the word ⟨ana/anaa⟩which was dominantly used in its short form in the private chat data is used more in its longer form in the Twitter data.

vowels less) because their intended audience is wider. This could also be the reason why we see more use of the longer forms of some words in the Twitter data. This might be one explanation why we find the mixed effects logistic regression analysis significant for the test data from the LDC but not from Twitter.

Choosing to drop the vowels when they seem essential to lexical disambiguation poses a question: How informative are the phonemes in Arabizi? Future work can potentially calculate phoneme informativity by training an n-gram model at the character level and obtaining the surprisal rates of every phoneme in the language given the previous phonemes. This might tell us that vowels are less predictable after some phonemes, which makes speakers more likely to spell them rather than dropping them. On the other hand, speakers might tend to drop vowels when they are more predictable given the phonemes they follow (Bell et al. (2009); Seyfarth (2014)).

One caveat in the current paper is that the most frequent words in the data set were mostly function words that include prepositions and pronouns, for instance the pronoun ⟨ana/anaa⟩ 'I' or the word ⟨tmam/tamam⟩ 'perfect'. This limited our choice of determining the short/long word pairs since many of these function words do not appear with preceding context in the data. For example pronouns like ⟨ana/anaa⟩ can occur as the first word in a sentence, or the word ⟨tmam/tamam⟩ would be sufficient on its own in a chat message. This could be overcome by considering a different set of short/long pairs of words from a larger data set of naturalistic speech. Overall, we are working with text data that are incentivised to be short (i.e., SMS messages and tweets), so we want to construct a list of content words that could provide more intuition into how vowel dropping is influenced by efficiency. Moreover, we constructed a list of word pairs for simplicity, although this list can be extended to include more possible spellings for each of the words chosen.

References

Akbar, R. (2019). Arabizi among kuwaiti youths: Reshaping the standard arabic orthography. *International Journal of English Linguistics*, *9*(1), 301.

Baert, G., Gahbiche, S., Gadek, G., & Pauchet, A. (2020, December). Arabizi language models for sentiment analysis. In *Proceedings of the 28th international conference on computational linguistics* (pp. 592–603). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.51 doi: 10.18653/v1/2020.coling-main.51

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, *60*(1), 92–111.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., ... Rambow, O. (2014). Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the emnlp 2014 workshop on arabic natural language processing (anlp)* (pp. 93–103).

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., ... Oflazer, K. (2018, May). The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1535

Darwish, K. (2013). Arabizi detection and conversion to arabic. *arXiv preprint arXiv:1306.6755*.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*.

Gordon, C. (2011). From speech to screen: The orthography of colloquial arabic in electronically-mediated communication.

Jaeger, T. F., & Tily, H. J. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 323–335.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR*.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition*, *126*, 313–318.

May, J., Benjira, Y., & Echihabi, A. (2014). An arabizi-english social media statistical machine translation system. In *Proceedings of the 11th conference of the association for machine translation in the americas* (pp. 329–341).

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Pimentel, T., Maudslay, R. H., Blasi, D., & Cotterell, R. (2020). Speakers fill lexical semantic gaps with context. *arXiv preprint arXiv:2010.02172*.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Shazal, A., Usman, A., & Habash, N. (2020). A unified model for arabizi detection and transliteration using sequence-to-sequence models. In *Proceedings of the fifth arabic natural language processing workshop* (pp. 167–177).

Sullivan, N. (2017). *Writing arabizi: Orthographic variation in romanized lebanese arabic on twitter* (Unpublished doctoral dissertation).

Tobaili, T. (2016). Arabizi identification in twitter data. In *Proceedings of the acl 2016 student research workshop* (pp. 51–57).

Zipf, G. K. (1936). *The psychobiology of language*. London: Routledge.