Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words

Gwilliams, L.[1] & Marantz, A.[1,2]

[1]*NYUAD Institute, New York University, Abu Dhabi*
[2]*Departments of Psychology and Linguistics, New York University*

**Address correspondence to*
Laura Gwilliams
New York University Abu Dhabi
Saadiyat Island
Abu Dhabi
UAE
PO Box 129188
Tel.: +971-563190446
E-mail: laura.gwilliams@nyu.edu

Additional contacts
Alec Marantz
New York University
10 Washington Place
New York, NY 10003
Tel.: (212) 9983593
E-mail: marantz@nyu.edu

Word count: 11652
Figure count: 5
Table count: 2

Running Header: *The influence of morphological structure on spoken word recognition*

Abstract

Although the significance of morphological structure is established in visual word processing, its role in auditory processing remains unclear. Using magnetoencephalography we probe the significance of the root morpheme for spoken Arabic words with two experimental manipulations. First we compare a model of auditory processing that calculates probable lexical outcomes based on whole-word competitors, versus a model that only considers the root as relevant to lexical identification. Second, we assess violations to the root-specific Obligatory Contour Principle (OCP), which disallows root-initial consonant gemination. Our results show root prediction to significantly correlate with neural activity in superior temporal regions, independent of predictions based on whole-word competitors. Furthermore, words that violated the OCP constraint were significantly easier to dismiss as valid words than probability-matched counterparts. The findings suggest that lexical auditory processing is dependent upon morphological structure, and that the root forms a principal unit through which spoken words are recognised.

Keywords: MEG; morphology; spoken word recognition; roots; decomposition; prediction; surprisal; Obligatory Contour Principle; Arabic

1. Introduction

1.1 Routes to word recognition

In modelling the structure of the mental lexicon, one of the most prevalent questions is the role that morphology plays in the organisation, production and comprehension of words. Historically the debate has been between "decompositional" and "whole word" theories of word recognition, with evidence over the past decade supporting a morphologically sensitive, decompositional approach in the visual modality of lexical processing. Behavioural masked priming studies for example, which have somewhat dominated the field of enquiry, have found consistent evidence for the decomposition of words with regular suffixation and pseudo-suffixation (e.g., teacher-TEACH; corner-CORN; Rastle, Davis and New, 2004; see Rastle and Davis, 2008 for a review). Corresponding results have also been established in the neurophysiological literature, supporting decomposition of regularly derived (e.g., Solomyak and Marantz, 2010) irregularly derived (e.g., Stockall and Marantz, 2006) and pseudo-suffixed forms (e.g., Lewis, Solomyak and Marantz; Whiting, Shtyrov and Marslen-Wilson, 2014). This body of research indicates that comprehending a visual word entails decomposition into constituent morphemes, which are linked to abstract representations in the lexicon for processing.

The influence of word-internal structure in spoken word recognition has been explored to a much lesser extent, and contention remains regarding the role of morphology in auditory processing. Methodologies for exploring the decomposition of complex words into morphemes during spoken word recognition include cross-modal priming, whereby an individual is presented with a masked visual word and asked to make a lexical decision on an auditorily presented target. Evidence from this paradigm appears to coincide with evidence from the visual domain of processing, whereby the root of a regularly derived complex word (e.g., government-GOVERN; Marslen-Wilson, Tyler, Waksler and Older, 1994; Kielar and Joanisse, 2010) or suffixed non-word (e.g., rapidifier-RAPID; Meunier and Longtin, 2005) is primed for recognition. Responses to morphological violations such as the incorrect use of verbal inflection have also been evidenced to elicit specific ERP response components, independent from semantic or syntactic lexical errors (Friederici, Pfeifer and Hahne, 1993). Furthermore, compound words that consist of two free stems (e.g., teacup) also appear to be decomposed into their constituents and incrementally integrated (Koester, Holle and Gunter, 2009), aided by prosodic information (Koester, 2014).

Two main theories of spoken word recognition can currently be recognised. A “continuous”, non-decompositional approach supports a strictly linear and morphologically insensitive method of auditory processing: The Shortlist B model as proposed by Norris and McQueen (2008) posits that auditory word recognition is based on the probability distribution of acoustic signals over time, whereby the likelihood of each incoming phoneme is predicted based upon all prior phoneme(s) that have been processed, regardless of word-internal structure. This theory is considered a full listing model as it assumes a lexicon that is structured in terms of whole word units rather than morphological constituents, in accordance with Butterworth (1983) and Janssen, Bi and Caramazza (2008). A recurring and prevalent objection to such a theory, however, is the necessary redundancy that would be caused by holding separate entries in the lexicon for morphologically related words such as “cover”, “uncover” and “covering”, for example (Wurm, 1997); although some suggest that using storage size as a measure of efficiency is misguided given the capacity of the human brain (Bybee, 1988; Sandra, 1994). In addition, from a linguist’s perspective, full listing models are not obviously compatible with the results of linguistic morphology (see Marantz 2013).

The “dis-continuous”, decompositional group of models holds a contrastive view. These theories support a morphologically structured lexicon and therefore a morphologically centred mechanism of auditory processing. From this perspective morphologically complex words are decomposed during word recognition, production and storage, and representations are formed on the basis of morphological constituents rather than whole words. By implication, a dis-continuous model of auditory processing would work on the basis of *morphological* recognition rather than whole word recognition. Consequently, each subsequent phoneme in the input is compared to possible morphemes and morphological continuations.

Experimental work has considered the uniqueness point (UP) to be an important factor in adjudicating between these two routes of auditory word recognition. The classic definition of UP refers to the point at which the word deviates from all onset-aligned words apart from inflectionally suffixed words and compounds, and has been shown to be an important determiner of lexical decision reaction-time (Marslen-Wilson and Welsh, 1978). This measure of UP assumes a continuous model of auditory word recognition, in agreement with Shortlist B, as it posits that the multimorphemic status of a cohort competitor formed through the affixation of derivational morphemes, whether or not these morphemes are productive in a language, is irrelevant to lexical recognition, with derived and un-derived forms treated equivalently. Recently, morphologically sensitive measures of UP have also been defined and positively assessed as predictors of lexical processing. For example, Balling and Baayen (2012) define the *complex uniqueness point* (CUP) as the point at which a suffixed word becomes uniquely distinguishable from all words that share the same stem, therefore considering derived morphological continuations as (morphological) competitors during recognition. Wurm (1997) focuses on the importance of prefixes to spoken word recognition and formulates the *conditional root uniqueness point* (CRUP) as the uniqueness point of the root given a particular prefix. Both the CUP and the CRUP were found to contribute significant predictive value to models of auditory (Wurm, 1997) and visual (Balling and Baayen, 2012) lexical decision tasks, in addition to the classic UP measure. Both authors therefore suggest that a combination of full-form processing and decomposition are involved in word recognition. Although these calculations do not constitute a processing model in their own right, they indicate that morphological structure is relevant to word recognition and motivate the formulation of a morphologically sensitive model of lexical processing.

1.2 Neuroimaging research of phoneme processing and prediction

Neurophysiological investigations into spoken word recognition suggest that the superior temporal gyrus (STG) is responsible for both low- and high-order processing of speech (Obleser and Eisner, 2008; Scott, Blank, Rosen and Wise, 2000; Scott and Johnsrude, 2003). A recent study (Mesgarani, Cheung, Johnson and Chang, 2014) investigated the role of the STG in processing acoustic information such as phonetic features, in order to establish how phoneme distinction arises during processing. Participants listened to 500 sentences of natural speech samples across a range of 400 native English speakers, and neurophysiological responses were recorded at the onset of each phoneme using direct inter-cranial recordings from the cortical surface of the STG. Distinct neural responses were found for phonemes differing on certain feature dimensions, such as manner of articulation for consonants (e.g., plosive vs. fricative), or the place and manner of articulation of vowels (e.g., low-back, high-front or glide); consistent responses were established across phonemes with shared features, regardless of the physical difference in acoustic realisation as a consequence of speaker differences. The neural populations recorded were found to be sensitive to phonetic features within the time-window of 150-200 ms post-phoneme onset; suggesting that the STG is responsible for low-level (but "abstract" categorical) processing of speech during this time course of activation.

Later in the time-course, the STG has also been associated with high-level processes such as the encoding of phonological prediction based on lexical knowledge. Gagnepain, Henson and Davis (2012) conducted a study that compared responses of learned novel words (e.g., *formubo*) as compared to existing similar words (e.g., *formula*) and baseline words to which the participants had no prior exposure (e.g., *formuty*). The learned novel word "formubo" served to delay the UP of "formula" until the final consonant, thus modifying the possible phonemes that could be predicted at "formu" and allowing for an assessment of segment prediction at the following phoneme. The authors used magnetoencephalography (MEG) to measure neurophysiological responses to experimental items pre- and post-UP (e.g., before and after the "l" in "formula") in order to assess how the trained novel words modified phoneme prediction. When comparing learnt and existing items, sensor-space analysis of the root mean square (RMS) of left-temporal MEG gradiometers found a reliable temporal cluster 280-350 ms after the onset of the UP; more activity was elicited for the novel over the existing words, suggesting that activity negatively correlated with the predictability of the divergent phoneme. No differences were observed pre-UP, also in accordance to theories of segmental prediction, as all information prior to the divergent phoneme supports both the existing and learnt lexical items. Source reconstruction of these neural responses localised the effect to the STG. In a model proposed by the authors they suggest that the STG is responsible for establishing a set of co-activated lexical candidates given the sensory input, in order to form competing hypotheses about which phonemes will be heard next. With each additional speech segment, any competitors that become incongruent with the input are eliminated, and the remaining cohort receive increased activation as likely lexical targets. If the materialised phoneme sequence does not match the expectations formed by possible outcomes, the resultant activity reflects an “error prediction signal”. This model therefore places competitors for word recognition at the forefront of segmental prediction, and the STG as the focal location for encoding responses related to segment prediction.

In order to investigate whether morphological structure enhances phoneme prediction, and therefore the effects of prediction error, Ettinger, Linzen and Marantz (2014) crossed morphological complexity with probability of word-final syllable during spoken word recognition, using MEG to measure responses at word offset to bi-morphemic and mono-morphemic disyllabic words (i.e., *swiftness* vs. *compost*) where each set of words included more probable and less probable second syllables. Activity localised to the STG and transverse temporal gyrus (TTG) displayed a main effect of second-syllable surprisal, and an interaction between surprisal and morphological complexity, whereby morphologically complex words showed an enhanced “surprisal” effect (where surprisal as a variable is a particular function of conditional probability rather than a psychological effect of “surprise” in the layman’s sense). This response was indexed by greater signal amplitude in a time window of 0-200 ms post word offset. The authors’ findings suggest that morphological structure bolsters phoneme prediction, leading to stronger prediction-error signal in the areas surrounding the auditory cortex. This finding is in corroboration to Wurm (1997) and Balling and Baayen’s (2012) results, suggesting that the neural mechanisms underpinning lexical recognition are sensitive to the shifts in probability distributions of upcoming suffix units, in the context of stem morphemes.

1.3 Spoken word recognition in Semitic languages

Phonological prediction, and its interaction with morphological complexity, has only been investigated in languages where both the phonemes and morphemes are linearly structured. Semitic languages offer an interesting case in this regard, as their internal structure allows for constituent morphemes to be organised in a non-concatenative manner (although concatenative affixation also occurs). Accounts of Semitic morphology support the existence of two primary structural units in the formation of open class words: the consonantal "root" (e.g., {$k_1t_2b_3$}) which holds the bulk of the "encyclopedic" semantic information, and the "pattern" (e.g., {$C_1aC_2aC_3$} ("C" = consonant) which conveys syntactic information. A whole word (in this case, *katab*[1]) is created when the consonants of the bound root slot into the relevant positions in the pattern (Doron, 2003). Although the specific morphological status of the pattern is under dispute within linguistics; for example, there is debate about whether the pattern is composed of two morphemes: the "vocalism" (e.g., {a-a}) and the skeletal "template" (e.g., $C_1VC_2VC_3$; "V" = vowel; see McCarthy, 1979), general consensus exists that the root forms a discrete morphological unit. The root morpheme in Arabic is similar to the root in Indo-European languages[2] (e.g., [appear] in the complex form [disappears]) in that it specifies a semantic field and forms the base of all morphologically related words across syntactic categories (Habash, 2010:43). For example, the tri-consonantal root {*ktb*} is also included in derived forms: [**k**aa**t**a**b**-a] *corresponded*; [**k**u**t**i**b**-a] *was written*; [**k**i**t**aa**b**] *book*; [**k**u**t**u**b**] *books*; [**k**u**tt**aa**b**] *writers* (Ryding, 2005:46); forming each word by placing the root into different patterns (i.e., *CaaCaC-a*, *CuCiC-a*, *CiCaaC*, *CuCub*, *CuCCaaC* respectively).

---

[1] All examples provided in Latin script follow Buckwalter's transliteration scheme

[2] In the current discussion, we take the "root" to be the base of a morphologically complex word once both inflectional and derivational morphology has been stripped away (e.g., *touch* from untouchables); the "stem" is the base once inflectional morphology alone has been stripped away (e.g., *untouchable* from untouchables).

Studies of Semitic languages have established an important role of morphology in word recognition. In visual masked-priming studies, evidence has supported morphological decomposition as indexed by shorter reaction times when prime and target share the same root (Arabic: Boudelaa and Marslen-Wilson, 2005; Maltese: Twist, 2006; Hebrew: Frost, Forster and Deutsch, 1997); similar findings have also been established in cross-modal priming studies, whereby words presented auditorily will prime recognition of visually presented words (Hebrew: Frost, Deutsch, Gilboa, Tannenbaum and Marslen-Wilson, 2000; Arabic: Boudelaa and Marslen-Wilson, 2001), in addition to case studies of root-specific aphasic speech errors (Prunet, Béland and Idrissi, 2000). Although less investigation has been conducted where the critical items were spoken words, the results of experiments in the auditory domain support similar conclusions. Mimouni, Kehayia and Jerama (1998) conducted an auditory morphological priming study of singular and plural nouns in Algerian Arabic and found evidence for root access through the decomposition of suffix-inflected and singular forms. Furthermore, Schluter (2013) found evidence for decomposition into root morphemes in a subliminal speech-priming paradigm where both the prime and target were auditory stimuli. The body of evidence therefore suggests that, similar to Indo-European languages, morphologically complex words are decomposed into their constituents units during recognition, supporting a dis-continuous model. However, there is a significant imbalance between evidence for decomposition in Semitic languages in visual processing and auditory processing, and the research thus far has been dominated by priming studies.

1.4 Obligatory Contour Principle

The existence of a root-specific constraint, the Obligatory Contour Principle (OCP: McCarthy, 1979), has also been used as evidence for decomposition and the abstract representation of consonantal roots. In Semitic languages such as Arabic and Hebrew, the OCP restricts the co-occurrence of two homorganic (OCP-Place) or identical consonants in a row within a root (e.g., [*XXY][3]). The constraint is crucially root specific, as consonant

[3] Occurrence of two homorganic consonants at the end of the root (e.g., [XYY]) also violates the OCP. However, the literature on Semitic languages argues that a bi-literal root [XY] can surface as [XYY] through spreading or copying of the second root consonant to fill out a prosodic template.

repetition is valid between a prefix and the first consonant of a root (i.e., X-XYZ). McCarthy explains this constraint within the framework of autosegmental phonology, suggesting that root consonants are represented in a separate autosegmental tier, and this tier is restricted by the OCP. The constraint is therefore assumed to act specifically upon the abstract representation of the root morpheme.

Recent studies have been conducted to determine the psychological reality of such phonotactic constraints by assessing whether they are formed from accidental patterns or linguistically significant generalisations. This can be determined by measuring the productivity of a given constraint for the language user, and the extent to which it informs phonological adaptation of foreign words. Becker, Ketrez and Nevins (2011) tested the former by conducting an experimental “wug test” of statistical regularities in Turkish, assessing the productivity of the phonotactic morphological constraint of laryngeal alternations. Their results allowed for precise discrimination between generalisations that are accidental and those that are phonologically motivated, and suggest that although a range of statistical regularities arises in language, only those that are grammatically motivated are used productively by speakers.

Similar experimental investigations have also been carried out in Arabic by asking individuals to rate nonce words for their “word-likeness”. Frisch and Zawaydeh (2001) investigated whether the OCP-Place constraint arises via analogy to possible words in the lexicon or through accessing an abstract phonotactic grammar. They found that the constraint effects in Arabic were graded in nature, whereby the more similar the two initial consonants, the less word-like the rating, suggesting that the phonotactic OCP-Place is more sophisticated than a symbolic description of possible words. Furthermore, there was no effect of neighbourhood density, which sets up the most likely competitors for recognition, suggesting that analogy to existing words was not the factor driving the perceived acceptability of a word’s structure.

In a later study by Frisch, Pierrehumbert and Broe (2004), the authors establish additional evidence that the strength of the constraint is graded relative to the similarity between the two initial consonants, and discuss related psycholinguistic evidence for the existence of similar constraints in other languages; for example, a number of Indo-European languages disallow root morphemes that are formed from the repetition of consonants in a $C_1C_2VC_2$ structure (e.g., English has *speak, smell* and *plate*, but not **speap, *smemm,* and **plale*; Domahs, Kehrein, Knaus, Wiese and Schlesewsky, 2009). They suggest that the cross-linguistic evidence for similarity avoidance may be an indication that repetition within speech processing is generally eschewed, placing the constraint in the domain of universal human cognition.

The OCP constraint is similarly manifest in Hebrew, a language that holds a comparable Semitic morphological structure to Arabic. Berent and Shimron (1997) investigated gemination of root-initial consonants in Hebrew, and found that nonce items containing the OCP constraint were rated as the least acceptable as words, when compared to root-final gemination and no gemination controls. In a later study, Berent, Vaknin and Marcus (2007) assessed the significance of root gemination (ssk, skk) when occurring in nominal paradigms that differed in the degree to which they support consonant repetition (CéCeC versus CiCúC). The authors found a significant interaction between the two units, whereby identical roots evoked significantly different acceptability ratings depending upon the constraints of the nominal paradigm. This result was interpreted as supporting the representations of stems rather than roots; however, it does not rule out the hypothesis that both units are represented separately but language users are sensitive to the co-occurrence of certain roots with certain patterns.

The results across both Semitic languages appear to support that language users are sensitive to the rules of root formation, and that these rules are based upon sub-lexical linguistic knowledge and not analogy to other existing words. Furthermore, it is proposed that the strength of the OCP within the Semitic root is due to the close proximity of the consonants when they are stored at a distinct level of lexical representation (i.e., in the "root-tier" as proposed by McCarthy [1979]). Such findings support abstract representation of the root in order to account for linguistic generalisation, and decompositional access to the root morpheme in order for sensitivity to the constraint to arise.

1.5 Predictions

Our study investigates the role of morphological structure in auditory word recognition through two routes of exploration. In using MEG to track neurophysiological responses to spoken words, we aim to explore the influence of morphological structure on phoneme prediction in the Semitic language, Modern Standard Arabic (MSA). By manipulating the degree of predictive power a root-final consonant has in a word, based either on all preceding sounds (a linear measure of prediction) or only root-internal sounds (a morphological and non-linear measure of prediction), it provides a comparison between whole word and root surprisal (Hale, 2001). If auditory processing of complex words requires access to morphological constituents in Semitic languages, it would be expected that phoneme prediction would be sensitive to morphological structure: When predicting the root-final consonant [t] in the root {*nbt*} only the consonants [nb] need to be considered in the predictive calculation. If, however, the whole word were recognised in a morphologically insensitive manner, all phonemes would be taken into consideration during lexical prediction: The probability of [t] within the word [*nabata*], *grow,* would be calculated based on all preceding phonemes [*naba*]. Alternatively, finding both surprisal calculations to be significant predictors of neural activity would suggest that access to whole word representations either occurs in parallel to, or as a consequence of, the processing of constituent morphemes. Critically, however, this would still support a system of word recognition that is sensitive to sub-lexical structure.

In accordance to previous studies (Ettinger et al., 2014; Gagnepain et al., 2012; Mesgarani et al., 2014) we expect to observe the neurophysiological effects of phoneme predictability in the STG and TTG, which will serve as our two regions of interest (ROIs) for the analysis. As the STG has been associated both with phoneme feature integration and segmental prediction within different points in the time-course of word recognition, we analyse two time-windows of interest associated with each stage of processing. First, based on when studies converge regarding the time-course of surprisal effects[4], we choose to select the time-window of 150-350 ms post-phoneme onset. Activity in this time-window would be associated with abstract segmental prediction based upon competitors for lexical recognition. Second, we choose to analyse an earlier time-window associated with the recognition and integration of phonemic features (Mesgarani et al., 2014), between 100-200 ms post critical phoneme onset. We predict that, since forming expectations of the realised phoneme segment (e.g., /b/) involves creating expectations for the features that make up the phoneme (e.g., plosive, voiced, nasal…), there should be a correlation between the most relevant surprisal measure (root or word-based) and activity in the STG within the time-window associated with phonetic feature processing. Finding morphological surprisal to be a significant predictor of neural activity either in addition to or independent of linear surprisal would support the hypothesis that auditory processing of morphologically complex words requires access to the constituent morphemes, and is underpinned by a morphologically-sensitive predictive mechanism. Furthermore, it would suggest that the mental lexicon holds representations of morphological roots, and auditory word recognition in Arabic crucially involves recognition of these roots during processing.

The second aim of this study is to build upon previous work on the OCP in auditory processing, and to explore the neurophysiological responses to these root-specific constraints. By analysing the MEG signal as time-locked to the onset of the second consonant, we compare OCP root violations borne from consonant gemination (e.g., {*qqr}, *qaqara* [قَقَرَ]), zero probability roots with non-existing C_1C_2 sequences (“illegal”) (e.g., {*vdh}, *vadaha* [هَدَثَ]), and roots that become invalid at the third consonant (“legal”) (e.g., {*fDz}, *faDaza* [ضفَز]). This comparison allows for insight into whether the neural response to a phoneme violating the OCP root-constraint differs from a phoneme that (“accidentally”) realises a probability of zero. Finding a distinct neurophysiological response to the OCP

4 Gagnepain et al. (2012) analysed within a broader window of 100-500 ms, and found significant effects in the RMS over MEG gradiometers from between 280-350 ms. Ettinger et al. (2014) analysed a 0-200 ms time window from critical word offset, which, as our phonemes are around 150 ms in length, translates to approximately 150-350 ms post phoneme onset.

violation would suggest that there is a discrete representation of these invalidating geminations, therefore supporting a decomposition approach to lexical processing.

2. Method

2.1 Participants

Twenty-five native Arabic, right-handed adults took part in the study (8 females, mean age = 20.8, *SD = 5.4*). All were literate in MSA and were in the process of completing undergraduate studies in a university that was taught in MSA. The individuals included native dialects from UAE, Yemen, Sudan, Palestine, Jordan and Syria. All had normal hearing and were recruited either from the NYU Abu Dhabi, or UAEU community. Written informed consent was provided by all participants prior to the experiment.

2.2 Materials

2.2.1 Word selection

Stimuli were selected from Arabic Gigaword Third edition (Graff, 2007). The corpus includes over 5 billion tokens of written text taken from nine sources of Arabic newswire data. As written Arabic typically does not include disambiguating diacritic short vowels, the entire corpus was parsed with MADAMIRA software, producing a fully diacriticised output (Pasha, Al-Badrashiny, El kholy, Eskander, Diab, Habash, Pooleery, Rambow and Roth, 2014). From the parsed corpus we extracted all words that followed a CVCVCV structure, which formed a pool of possible stimuli items. We chose this pattern structure because it realises one of the more common patterns in Arabic and it ensured that the root morpheme was always completely discontinuous. Furthermore, although we were always analysing neural activity from the final consonant, we chose words that ended in a final vowel in order to stabilise phoneme quality across stimuli and avoid word-close effects. Long vowels were not included in the extracted words due to their difference in orthographic representation from short vowels. We did not define the Sukun diacritic marker (ـْ) as a vowel in the same way we defined other diacritics, as it is realises a placeholder for silence rather than a phonological representation produced in spoken Arabic. Due to this, none of the words we selected included this Sukun diacritic.

Linear and morphological surprisal measures were computed based on the realisation of the third consonant. In order to calculate this, frequency counts were extracted from the entire corpus for the CVCVC, C-C-C, CVCV and C-C structures that appeared in the potential stimuli pool. For example, for the stimuli item *nabata* we calculated the frequency of *nabat, n-b-t, naba* and *n-b* in the raw corpus, where dashed lines allowed for the realisation of any phoneme. Linear surprisal was calculated as the negative log of its conditional probability given all preceding phonemes, as taken from Hale (2001):

$$-\log_2 \left(\text{freq[CVCVC]} / \text{freq[CVCV]} \right) \qquad -\log_2 \left(\text{freq[NABAT]} / \text{freq[NABA]} \right)$$

Morphological surprisal was calculated as the negative log of the root's conditional probability given all preceding consonants:

$$-\log_2 \left(\text{freq[C-C-C]} / \text{freq[C-C]} \right) \qquad -\log_2 \left(\text{freq[N-B-T]} / \text{freq[N-B]} \right)$$

From our stimuli pool we placed words in four contrastive bins based on high and low values of morphological and linear surprisal to yield a total of 320 items. A word was considered to have "low surprisal" with a value less than 3, and "high surprisal" with a value greater than 5. Additionally, all items with a surprisal value of greater than 5 in one measure and less than 3 in the other had at least 3 log. values of difference between the morphological and linear measures. Average surprisal values across items are presented in Table 1.

(Table 1 about here)

Linear and morphological surprisal measures were de-correlated during stimuli selection (*r* = -0.0074) – a difficult but possible task. All 320 words were then rated for familiarity online using *Amazon Mechanical Turk* (see Buhrmester, Kwang and Gosling [2011] for a discussion of this tool). Words with the lowest familiarity were excluded from the final stimuli list, to yield a total of 280 words: 72.9% were verbs (perfective, third person, masculine, singular; e.g., *nabata,* grow); 17.2% were nouns (singular, reduced/construct state; e.g., *darari,* harm/injure); and 9.9% were ambiguous between the two word classes (e.g., *halaba,* milk). All words consisted of six phonemes in a CVCVCV structure.

2.2.2 Non-word selection

A total of 280 non-words were included in the experiment in order to form a 50% split between words and non-words. Potential items were selected by first creating a list of all CVCVCV structures that did not appear in the parsed corpus, using the same phoneme restrictions as the word selection. The non-word "uniqueness point" (i.e., the point in which the item can no longer become a word) was manipulated so that it occurred at the second consonant for 140 non-words, at the second vowel for 70 non-words and at the third consonant for 70 non-words. Words where the UP fell at the second consonant was further separated into two conditions: OCP violation and "illegal". The OCP violations were defined as words where the first root-consonant was identical to the root-second consonant (e.g., *qaqara* [قَقَرَ]). Illegal violations occurred when the second consonant never occurred after the first, but the two consonants were different (e.g., *vadaha* [هَدَثَ]).

The difference between words and non-words could not be determined from acoustic features alone: stimuli sets did not differ in length or complexity, and words and non-words were spoken in a random order during recording to avoid order effects. Five individuals with no knowledge of Arabic were asked to perform the same lexical decision experiment, and did not perform above chance level (*p* > .5).

2.2.3 Stimulus recording and phoneme marking

All words and non-words were recorded by a native Arabic speaker in a single session using an Neumann U87 Microphone and Avalon VT-737SP preamplifier. Each item was read three times, and the second production of the word was always selected to allow for consistent intonation across stimuli.

Critical phoneme onsets (second consonant for non-words, third consonant for words) were marked using Praat software (Boersma and Weinlick, 2009). Phoneme boundaries were identified manually for each item; these were clearly identifiable by eye for the fricative and plosive consonants, but were more difficult to identify in glide consonants – in this case, extra care had to be taken through the combination of both visual and auditory inspection of the spectrogram and formant transitions. An annotated spectrogram example is presented in Figure 1.

(Figure 1 about here)

2.3 Procedure

All participants' head shapes were digitised using a FastSCAN laser scanner to allow for source localisation and coregistration (Polhemus, VT, USA). Digital fiducial points were recorded at five points on the individual's head: the nasion, anterior of the left and right auditory canal, and three points on the forehead. We placed marker coils at the same five positions in order to localise that person's skull relative to the MEG sensors. These marker measurements were recorded both immediately prior and immediately after the experiment in order to correct for movement during the recording.

MEG data were recorded continuously using a 208-channel axial gradiometer system (Kanazawa Institute of Technology, Kanazawa, Japan), while participants lay in a dimly-lit magnetically shielded room. Data were recorded with a 1000 Hz sample rate and low-pass filtered on-line at 200 Hz.

The experiment consisted of an auditory lexical decision task. Stimuli were presented binaurally to participants though tube earphones (Aero Technologies), using Presentation stimulus delivery software (Neurobehavioral Systems). Each trial of the experiment consisted of a fixation-cross displayed for 200 ms, followed by the onset of the auditory stimulus. Two hundred milliseconds after the offset of the stimulus, the choices [كلمة لا] (*non-word)* (presented on the left) and [كلمة] (*word)* (presented on the right) appeared onscreen. Participants were given a response box with two adjacent buttons, and pressed the left button when they thought the item was a non-word, and the right button to indicate that they recognised the given stimulus as a valid word of Arabic. The short delay between word offset and response cue was selected in order to ensure that activity in our time-windows of analysis were not interrupted by a visual response, whose timing after critical consonant onset would be slightly different for each item. As the response cue and button correspondence was kept stable across the entire experiment, it is unlikely that the 200 ms delay would postpone initiation of lexical processing or motor planning of the decision, but raw frequency times should be interpreted in light of this 200 ms gap between word offset and initiation of response timing. Following a response, a fixation cross was displayed and remained on screen until the participant pressed a button to move forward. Participants were instructed to use this time to blink and produce any other muscular movements that they deemed necessary. The trials were organised into 8 blocks, providing a break between each block for the participant to rest. Stimuli order was randomised across blocks, and each participant received a different randomisation. Each experimental recording was conducted in a single session and lasted around 30 minutes.

2.4 Analysis

The procedures used for preprocessing first involved removing noise from the raw data by exploiting eight magnetometer reference channels located away from the participants' heads; using the Continuously Adjusted Least Squares Method (CALM; Adachi, Shimogawara, Higuchi, Haruta, & Ochiai, 2001), with MEG160 software (Yokohawa Electric Corporation and Eagle Technology Corporation, Tokyo, Japan). The noise-reduced MEG recording, the digitised head-shape and the sensor locations were then imported into MNE-Python (see Gramfort, Luessi, Larson, Engemann, Strohmeier, Brodbeck, Parkkonen and Hämäläinen, 2014). Data were epoched from 100 ms pre-stimulus onset to 1200 ms post-stimulus onset. Artifact rejection consisted of manual rejection of trials that contained blinks and other motor artifacts; on average this removed 11% of participants' trials (range: 56% to 0.7%).

Neuro-magnetic data were co-registered with the FreeSurfer average brain (CorTechs Labs Inc., Lajolla, CA), first by scaling the size of the average brain to fit the participant's head-shape, aligning the fiducial points, and conducting final manual adjustments to minimise the difference between the headshape and the FreeSurfer average skull. Next, an ico-4 source space was created, consisting of 2562 potential electrical sources per hemisphere. At each source, activity was computed for the forward solution with the Boundary Element Model (BEM) method, which provides an estimate of each MEG sensor's magnetic field in response to a current dipole at that source. The inverse solution was computed from the forward solution and the grand average activity across all trials. Data were converted into noise-normalised Dynamic Statistical Parameter Map (dSPM) units (see Dale, Liu, Fischl, Buckner, Belliveau, Lewine and Halgren, 2000), employing an SNR value of 1. The inverse solution was applied to each trial at every source, for each millisecond defined in the epoch, employing a fixed orientation of the dipole current that estimates the source normal to the cortical surface and retains dipole orientation. We defined our two ROIs (STG & TTG) based on previous studies (Ettinger et al., 2014; Gagnepain et al., 2012). The STG consisted of 83 vertices, and the TTG consisted of 24 vertices (see Figure 2 below). Activity in these regions was separately averaged for each millisecond in our epoch, to produce a time-source of activity for each trial for each subject.

(Figure 2 about here)

For all statistical tests, we conducted mixed-effect model analyses using the *lme4* package (Bates, Maechler & Bolker, 2012) in *R* (R Core Team 2012). For the experimental manipulation of the valid word items, each model included a by-item intercept, a by-subject intercept and by-subject slopes for all the independent variables (Morphological Surprisal, Linear Surprisal, Root Frequency, Surface Frequency, Familiarity, Word Order); as no interactions were predicted, they were not included in the model. The full model included fixed effects for all independent variables; this created a maximal random effect structure, following Barr, Levy, Scheepers and Tily (2013). In order to test each variable of interest, our two predictors (Linear Surprisal and Morphological Surprisal) were removed in turn from the fixed effects and compared to the full model. Random effects were maintained for all variables in all models. Analyses fit a linear mixed model to reaction times (RTs) and neural data, and a mixed logit model for the binomial accuracy response (see Jaeger, 2008).

For neural analyses a cluster permutation test was conducted following the same procedure as described in Solomyak and Marantz (2009:193): we computed the correlation coefficient of the mixed-effects model detailed above for each millisecond within our separate time-windows. Temporal clusters were identified when effects in consecutive time points exceeded the $t = 1.96$ significance threshold, and were subject to multiple comparison correction following Maris and Oostenveld (2007). The *p*-value of the largest significant cluster was computed based on the consecutive t-values that exceeded the critical threshold, as tested against 10,000 permutations. This allowed us to assess the most significant cluster within the given temporal window for each predictor and each ROI.

To analyse the non-word items, a mixed-effects model regression was conducted (again with the *lme4* package in *R*) using Condition (OCP violation [*XXY], Illegal [*XZY] and Legal bi-phoneme pair [XYZ]) and Presentation Order as fixed effects and by-subject slopes in the design. The model also included by-item and by-subject intercepts, using a linear mixed model for RT and neural data, and a logit model for accuracy. In order to determine differences between conditions, this full mixed model was assessed with generalised hypothesis testing with Tukey correction using the *glht* package in *R*. For the neural data this was calculated at each millisecond of our window of interest; the t-value computed at each time-point for the pairwise difference between conditions was used to form temporal clusters for analysis employing the same method as used for the word items.

3. Results

3.1 Word Manipulation

3.1.1 Behavioural

Reaction time (RT) and accuracy of responses to the valid Arabic words were analysed as dependent measures of our variables of interest. Trials with reaction times greater than 2.5 standard deviations from either the by-subject or by-item mean were removed from the final analysis, eliminating 1.4% of the trials.

When assessing the significance of each of our independent measures to the statistical models of the behavioural results, we found Morphological Surprisal to approach significance for RT (χ^2 = 3.71, t = 1.93, p = .053, whereby greater surprisal led to longer latencies), but not for accuracy (χ^2 = .04, t = 0.21, p = .83). Linear Surprisal was not a significant determiner for either measure (RT: χ^2 = .54, t = 0.74, p = .46; Accuracy: χ^2 = .009, t = 0.09, p = .924). Both behavioural measures showed significant main effects of familiarity (RT: χ^2 = 8.64, t = -2.9, p = .003; Accuracy: χ^2 = 45.7, t = 6.76, p < .001, whereby a more familiar word was faster and more accurately identified). Order of presentation was a significant determiner of RT (χ^2 = 68.75, t = -8.29, p < .001, indicating that responses got faster as the experiment progressed) but not of Accuracy (χ^2 = .95, t = -0.97, p = .33).

3.1.2 Familiarity Ratings

Correlational analyses between familiarity ratings and our variables of interest are displayed in Table 2 below. Significant correlations were observed between Morphological

Surprisal and Familiarity (r = -0.195, t = -2.99, *p* = .003); Morphological Surprisal and Root Frequency (r = -0.348, t = -8.56, *p* < .001); Linear Surprisal and Surface Frequency (r = -0.261, t = -4.41, *p* < .001); Linear Surprisal and Root Frequency (r = -0.124, t = -2.3, *p* = .02); Surface Frequency and Familiarity (r = 0.214, t = 3.65, *p* < .001), although it should perhaps be noted that the raw correlation coefficients were not particularly high. No other correlations reached significance (r's < .1).

(Table 2 about here)

3.1.3 Neural

Activity was averaged over the sources in each ROI and time-locked to the onset of the third consonant in each word. Prior literature motivated two time-windows of interest: 100-200 ms to evaluate phonetic feature analysis (Mesgarani et al., 2009); and 150-350 ms, which has been implicated in phoneme prediction in two previous studies (Ettinger et al., 2014; Gagnepain et al., 2012). The cluster permutation test followed the same procedure as described in Solomyak and Marantz (2009:193, see Section 2.4 above).

Figure 3 below displays the significance of our two measures of surprisal across time in our two ROIs. In the earlier time-window of interest, we observed a main effect of Morphological Surprisal between 130-156 ms (*p* = .048, whereby greater surprisal led to more [negative] activity) in the STG, but no clusters were formed in the TTG. Linear Surprisal was not significant within this time window in either ROI. For the later window of interest, associated with segmental prediction, the STG displayed a main effect of Morphological Surprisal between 289-342 ms (*p* = .023, again where more surprisal led to stronger [negative] activation). The largest cluster for Linear Surprisal was marginally significant between 277-306 ms (*p* = .079). In the TTG Morphological Surprisal showed a significant effect between 294-338 ms (p = .027, higher surprisal corresponding to stronger [positive] activity), and no clusters met the threshold (*t* > 1.96) for Linear Surprisal.

(Figure 3 about here)

3.2 Non-word manipulation

3.2.1 Behavioural

Condition was a significant predictor of RT (χ^2 = 4.96, t = 2.04, p = .029) but not Accuracy (χ^2 = 6.2, t = -0.1, p = .1). There was no main effect of Presentation Order for either measure (RT: χ^2 = .18, p = .86, Accuracy: χ^2 = .018, t = -0.001, p = .9). In order to determine which differences between conditions were driving this effect, we conducted planned pairwise t-tests with Tukey correction using the *glht* package in R: words in the OCP condition were responded to significantly faster than both Illegal (z = -3.22, p = .007) and Legal bi-phonemes (z = -5.85, p = .001), and Illegal was significantly faster than Legal bi-phonemes (z = -3.03, p = .013). This finding suggests a graded nature of response: the OCP violations were the easiest to reject as words, followed by Illegal followed by Legal bi-phoneme pairs (see Figure 4 below). It of course is not surprising that the Legal items were rejected as words later than the other two conditions given that the point of divergence from a real word was later for these items.

(Figure 4 about here)

3.2.2 Neural

Activity in each ROI was time-locked to the onset of the second consonant in each word, and permuted over the same 150-350 ms time-window as the word analysis. The cluster permutation test followed an identical procedure as the word manipulation, although this time using generalised linear hypothesis testing to compare each condition in turn (Ilegal vs. Legal; Illegal vs. OCP; OCP vs. Legal). The t-values for these comparisons across time are displayed in Figure 5 below.

(Figure 5 about here)

In the STG, no clusters were significant (p's > .1). In the TTG, significant differences were found between Illegal and Legal conditions at 210-260 ms (p = .014). We also observed differences between OCP and Legal at 242-286 ms (p = .024). The comparison between Illegal and OCP conditions was not significant, but the largest cluster was formed at 208-223 ms (p = .14).

4. Discussion

The aim of the present study was to investigate whether spoken words that conform to a root and pattern morphological structure are processed through morphological units or (only) as whole word items. To address this question we probed the status of the root morpheme using two experimental manipulations. First we contrasted measures of phonological prediction of the root-final consonant, based either on the root morpheme alone (i.e., negative log of the conditional probability that "b" occurs after "kt" - Morphological Surprisal) or on all preceding phonemes (i.e., negative log of the conditional probability that "b" occurs after "kata" - Linear Surprisal). This opposes lexical identification of words based on morphological competitors (decomposition theory, which implicates non-linear processing) to lexical identification based on whole word competitors (continuous theory, which implicates linear processing). The second manipulation investigated the root-specific constraint, the OCP. We compared items where the second root-consonant was identical to the first (an OCP violation), to items where the second consonant never occurred after the first but realised a different phoneme. This allowed us to test the significance of this root constraint in an environment of identical probability values.

4.1 Effects of surprisal

Relative to the first manipulation, we found surprisal to be a significant determiner of neurophysiological activity in both the STG and TTG, whereby more activity was elicited for less predictable phonemes. This effect was observed in two time-windows: between 130-160 ms, and later between 280-340 ms from critical phoneme onset. The time-course is important to understanding the neural mechanisms supporting phoneme surprisal in our results, as different latencies are associated with different stages of linguistic processing.

The later surprisal effect corresponds to evidence that activity in the STG between 250-450 ms reflects an N400m "lexical-semantic analysis" stage of auditory processing (see Salmelin, 2007), associated with high-level processing of speech. This finding closely converges with previous studies (Ettinger et al., 2014; Gagnepain et al., 2012), which also found phoneme surprisal to correlate with activity in these regions at a similar time. Gagnepain and colleagues propose that predictions of upcoming phonemes are built from co-activated lexical candidates, which compete for recognition by making incompatible hypotheses for which sound will be heard next. Activity therefore reflects a mismatch response between the most expected phoneme sequences and the sensory input. Crucially, this interpretation positions lexical candidates as drivers of prediction, whereby a "candidate" is assumed to be any lexical unit that is represented in (and can therefore be extracted from) the lexicon. From this perspective, effects of Morphological Surprisal would be the consequence of predictions formed from potential root candidates, and effects of Linear Surprisal would be the consequence of predictions formed from potential whole word candidates.

Our results show that Morphological Surprisal was a significant determiner of neural activity within the time-course and regions associated with an error prediction signal. As this measure of surprisal calculates probability based on root competitors independent of whole word competitors, it suggests that root morphemes may be lexically represented units that are used to form predictions of upcoming phonemes and target words. Such a position is consistent with decomposition models of spoken word recognition, which suggest that the root morpheme is the central unit through which whole words are organised and processed. For example, prior research has found evidence for root access in Semitic languages (Hebrew: Frost et al., 2000; Arabic: Boudelaa and Marslen-Wilson, 2001), and comparative findings have also been established in Indo-European languages, whereby masked visual presentation of the root morpheme (i.e., GOVERN) aids overt auditory recognition of the complex form (i.e., government) (behaviourally: Marslen-Wilson, et al., 1994; with EEG: Kielar and Joanisse, 2011), and overt auditory priming facilitates recognition of a morphologically-related auditory target, absent of semantic overlap (e.g., SUBMIT, permit; Emmorey, 1989). These results have been interpreted as supporting decomposition and access to a representation of the root morpheme during spoken word recognition.

Recent work within the theoretical framework of Distributed Morphology (see Halle and Marantz, 1993) proposes that roots, uncategorised for syntactic category, form the universal building block of open class words. Therefore, finding that a mechanism of root recognition may be relevant to spoken Arabic is particularly remarkable given that, unlike Indo-European languages, the auditory input of the root is a discontinuous set of sounds that cannot be neatly separated from the speech stream, nor pronounced in isolation. Our results support the existence of a mechanism that is able to extract each component (in this case, consonant) of the root morpheme from the whole word, and set up a comparison between 1) the sensory evidence and 2) possible realisations of the root and their relative likelihood of occurrence. Incoming phonemes would presumably have to be separated into morphemic categories as they materialise over the speech stream, and recognised relative to mental representations of possible roots. This suggests that it is not required of morphemic representations to be pronounceable, nor to have stand-alone semantic meaning (i.e., the root {ktb} does not have specific meaning or pronounceability until combined with a pattern). Such a finding is therefore not only relevant to Semitic languages, but also informs our more global understanding of lexical representations consistent with a decompositional theory of such as Distributed Morphology.

The earlier surprisal effect we observe in the STG corresponds to a time-course associated with low-level processing; it has been proposed that acoustic-phonetic analysis of a sound form occurs from around 100 ms post-phoneme onset, and the mechanisms at work involve pre-lexical processes such as phonological identification. Findings from direct multi-electrode cortical surface recordings, for example, have established 150 ms post-phoneme onset as the time-point in which the STG is sensitive to the classification of phonetic features from the auditory input (Mesgarani et al., 2014). Furthermore, Uusvuori, Parviainen, Inkinen and Salmelin (2008) conducted an MEG study that identified an N100m response within a 75-160 ms time-window to phonologically unexpected (but semantically valid) words in list context (e.g., salad, samba, satin, *river*), which also corresponds to phonological analysis prior to lexical selection.

Morphological Surprisal was the only variable found to significantly correlate with activity in the STG during this earlier time-window, from 130-156 ms. It would be reasonable to suggest that in the process of building expectations of upcoming lexemes, and the phonemes through which they materialise, predictions of the phonetic features of particular phonemes also develop, yielding an error signal for a mismatch in this time period. Our results only support the STG as involved in both of these stages of processing; activity in the TTG did not correlate with our variables until the time-window associated with segmental surprisal.

In addition to the morphological effect, we also found a marginal effect of Linear Surprisal in the STG during the time-window associated with segmental prediction, between 250-350 ms, but not the time-window associated with phoneme feature identification. It should be noted that these effects are statistically independent; our materials were chosen to orthogonalise Linear and Morphological surprisal, and our statistical modelling confirmed the significance of each measure independent of the other. Although the whole word surprisal predictor was not as strong as the measure of root surprisal, this experiment does not rule out that whole word competitors may also be relevant to spoken word recognition, and methodological explanations for the weakness of the linear effect we observed should be considered when interpreting this result. For example, it is possible that the disambiguating MADAMIRA parser contained inaccuracies when retrieving diacritic (vowel) information, thus maintaining a larger margin of error in the Linear Surprisal calculations as compared to the Morphological Surprisal calculations. This would therefore reduce the strength of any correlation with neural activity that we observed.

Whole word surprisal effects, in the context of our results, are compatible both with fully decompositional and hybrid models of spoken word recognition. From a decompositional perspective, the frequency of morphological constituents, and the frequency with which they combine, may feed representations of morphologically complex units. Such a standpoint entails obligatory non-linear processing of input phonemes, driven by the morphological structure. Neurophysiological correlation with Linear Surprisal from this view would reflect a parasitic effect from the frequency with which morphological representations are accessed and combined by the listener (see Marantz, 2013). A contrary interpretation of this finding calls upon the number of hybrid models that have recently been proposed (see Baayen, 2014 for a review). These theories posit that both decomposed and whole word processing are possible routes to recognition, and therefore suggest that lexical processing does not necessarily involve recognition through decomposition. From this standpoint, both linear and non-linear processing mechanisms are possible and supported, and which prevails is determined by what method is the most efficient for a given word. Although, as discussed above, there are a number of possible explanations for the weaker linear correlation, finding that surprisal calculated from the root elicited a response earlier than surprisal calculated from the whole-word is consistent with the assumptions of Distributed Morphology (in that it is consistent with root access prior to whole word access), but not those of hybrid "race" models, which posit that morphemic and surface forms are integrated in parallel. The observed early asymmetry between Morphological and Linear surprisal measures therefore provides a viable adjudication between models that do and do not assume obligatory decomposition of morphologically complex words, and suggest that the phonemes of root and pattern Arabic words are predicted in a non-linear fashion.

To conclude the discussion of the surprisal manipulation, our findings suggest that the neural mechanisms underpinning lexical recognition follow the processes of our morphologically sensitive model, whereby lexical access primarily entails root recognition. Due to the non-concatenative morphological structure of the words we selected, we found evidence that the processing and prediction of upcoming phonemes is not necessarily linear, but rather the anticipation of upcoming speech sounds is determined by the morphological structure of a given word. We propose that root-relevant phonemes are extracted from the speech stream and utilised to modify probability distributions over upcoming root phonemes. This position is inconsistent with continuous models of processing such as Shortlist B, as such a theory deems morphological structure as peripheral to word recognition. Instead these results converge with decompositional mechanisms of spoken word recognition by placing the root morpheme as the central unit through which whole words are organised and processed.

4.2 Offline familiarity-ratings

In addition to the behavioural and neurophysiological results, we also found a significant negative correlation between Familiarity and Morphological Surprisal, whereby a morpheme with a less predictable root-final consonant was rated as less familiar, independent of any correlation between root surprisal and word frequency. The rating task presented whole words (including vowel diacritics) visually, suggesting that predictability of the root based on the first two consonants is relevant even when all of the word is available in the same instance. Interestingly there was no correlation between Linear Surprisal and Familiarity, suggesting that the familiarity rating is being driven by something relevant to the root morpheme.

Previous studies (Schreuder and Baayen, 1997; Baayen, Feldman and Schreuder, 2006), found linguistic variables such as surface frequency, summed inflectional frequency, written-spoken ratio and morphological connectivity to correlate with familiarity or "subjective frequency"; linguistic variables which also significantly predictor reaction time in lexical decision tasks. Our results also demonstrate parallels between RT and Familiarity as dependent measures, despite their differing modality. As previous studies support access to the Arabic root during reading (Boudelaa and Marslen-Wilson, 2005), it may be postulated that the likelihood of a third consonant of the root evokes a comparison with other root competitors, therefore deeming more likely roots as more familiar. The results of studies conducting relative frequency tasks, whereby participants decide which of two words or phrases are the most frequent, have shown that language users are sensitive to subtle differences between outcomes close in frequency (Shaoul, Westbury and Baayen, 2013). It may be suggested, therefore, that when these comparisons arise covertly, individuals remain sensitive to the relative likelihood of outcomes and use this information in order to inform subjective ratings.

Given the evidence to suggest that letter identification within words up to four letters in length occurs simultaneously (Adelman, Marquis and Sabatos-DeVito, 2010), it is possible that similar "surprisal effects" would be observed regardless of a phoneme's position within a morpheme. For example, varying the likelihood of the first consonant (e.g., k within *kataba*) may influence familiarity ratings to the same extent as our investigation into the final consonant. If a similar correlation were established for root-initial and -internal consonant surprisal, it would suggest that the important factor relative to familiarity is the likelihood of the root outcome given the other two consonants. This would suggest that "subjective frequency" is formed through drawing a comparison between competing root representations, and ease of identification of the root morpheme; supporting that the feeling of familiarity with a word is formed through processing the root morpheme, despite its necessarily abstract representation.

4.3 Root-specific constraint: OCP

The secondary aim of our study was to explore the OCP and its importance to the recognition of spoken Arabic words. Previous studies have established the psychological reality of this constraint and its influence upon lexical validity for native Arabic speakers, using this as evidence for independent representations of the root morpheme.

Our behavioural results showed that items containing an OCP violation were faster to be identified as non-words than items where a second non-identical consonant had a probability of zero given the first. This finding suggests that the OCP violation within the root [*XXY] provides a stronger cue for lexical invalidity than the invalid consonant pair alone [*XZY], where probability value was equal in both cases. These results converge with previous studies conducted in Arabic and in Hebrew. For example, Berent and Shimron (1997) used a non-word rating task to investigate responses to constraint violations, finding that non-words derived from roots with initial gemination were rated as the least word-like, as compared to non-words with root-final gemination and no gemination. This was also found for words with affixation and root-internal infixation, showing a robust effect regardless of the position of the root within the word. In line with our findings, the authors propose that sensitivity to root structure, and the specific location of consonant repetition, support the status of the root as a distinct morphemic unit. Importantly, finding root-specific results for nonce words provides evidence for morphological decomposition without the confound of semantic and orthographic overlap between whole words and the morphological constituents through which they are formed - a recurring counterargument for the existence of distinct morphological representation (Seidenberg, 1987). Rather, evidence that native Arabic speakers are sensitive to the root-specific phonological constraint within novels words suggests that any root, regardless of lexicality, must be represented separately from its word pattern.

Unlike the behavioural results, we did not observe a reliable difference between Illegal [*XZY] and OCP [*XXY] roots in the neurophysiological analysis. Rather, both regions displayed a trend towards greater activity for Illegal than OCP violations around 200 ms post second consonant onset, although these effects did not reach significance. The only other study known to the authors' utilising neurophysiological techniques to explore OCP constraints was conducted by Domahs et al. (2009) using EEG. They investigated OCP violations in German, comparing geminating $C_1C_2VC_2$ non-words, $C_1C_2VC_3$ pseudo-words and existing words. The behavioural difference between conditions was consistent with prior literature and our present results, whereby participants were faster to reject non-words containing an OCP violation than plausible pseudo-words. However, they also found post-lexical differences between pseudo-words and OCP violating conditions in the EEG analysis, in a time-window after the completion of phonological processing.

There are a number of possible explanations for our null neurophysiological result in the context of robust behavioural effects. Firstly, there is evidence to suggest that physical co-occurrence of the same phoneme can reduce neural responses to the repeated sound in the STG and primary auditory cortex (Bergerbest, Ghahremani and Gabrieli, 2006); therefore, any difference observed as a consequence of the OCP violation may be dampened by a repetition effect. Secondly, it has been suggested that the core of the OCP is avoidance of shared phonological features, rather than direct phoneme repetition, as some languages allow gemination but not homorganic articulation. In a rating task conducted by Frisch and Zawaydeh (2001) for example, OCP gemination effects were found to behave differently from the graded OCP-Place similarity effects, and did not conform to the correlation between shared features and lexicality rating. In order to fully assess the neural mechanisms underlying the consistent behavioural literature on OCP constraints, it may therefore be better for future studies to explore root-initial consonants that share a graded similarity of articulatory features, rather than direct consonant gemination. This would avoid possible dampening phoneme repetition effects, and allow for the analysis of neurophysiological responses of as a function OCP-Place.

5. Conclusion

The results of the present study address a number of questions regarding spoken word recognition. First, we replicate the finding that differences between segmental prediction and sensory outcome are coded neurologically in the STG and TTG, extending the investigation to languages with a root and pattern morphological structure. Second, our measure of morphological surprisal was found to be a significant determiner of neurophysiological response, familiarity rating and response latency, supporting a morpheme-centred model of auditory word processing in non-concatenative languages. Finally, we found that a specific morphological constraint, the OCP, appears to trigger responses that are separable from probability alone, supporting a distinct representation of the morphological root in the lexicon.

Acknowledgements

This research was supported by the NYU Abu Dhabi Research Council from NYUAD Institute, under Grant G1001 (AM, LG). We would like to thank Matthew Tucker and Kevin Schluter for guidance on stimuli creation and Arabic linguistic theory, Nizar Habash and

Ahmed El Kholy for software assistance with MADAMIRA, and Christian Brodbeck and Benoit Merchant for computational support. We are also very grateful for the suggestions provided by Adam Ussishkin on an earlier draft of this manuscript.

References

Adachi, Y., Shimogawara, M., Higuchi, M., Haruta, Y., & Ochiai, M. (2001). Reduction of non-periodic environmental magnetic noise in MEG measurement by continuously adjusted least squares method. *IEEE Transactions on Applied Superconductivity*, *11*(1), 669-672.

Adelman, J. S., Marquis, S. J., & Sabatos-DeVito, M. G. (2010). Letters in words are read simultaneously, not in left-to-right sequence. *Psychological Science*.

Baayen, R. H., Experimental and psycholinguistic approaches to studying derivation. Lieber, R. and Stekauer, P. (eds) , *Handbook of derivational morphology*. Oxford: Oxford University Press, in press.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290-313.

Balling, L. W., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*(1), 80-106.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes.*

Becker, M., Ketrez, N., & Nevins, A. (2011). The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language*, *87*(1), 84-125.

Berent, I., Vaknin, V., & Marcus, G. F. (2007). Roots, stems, and the universality of lexical representations: Evidence from Hebrew. *Cognition*, *104*(2), 254-286.

Berent, I., & Shimron, J. (1997). The representation of Hebrew words: Evidence from the obligatory contour principle. *Cognition*, *64*(1), 39-72.

Bergerbest, D., Ghahremani, D. G., & Gabrieli, J. D. (2004). Neural correlates of auditory repetition priming: reduced fMRI activation in the auditory cortex. *Journal of Cognitive Neuroscience*, *16*(6), 966-977.

Boudelaa, S., & Marslen-Wilson, W. D. (2005). Discontinuous morphology in time: Incremental masked priming in Arabic. *Language and Cognitive Processes*, *20*(1-2), 207-260.

Boudelaa, S., & Marslen-Wilson, W. D. (2001). Morphological units in the Arabic mental lexicon. *Cognition*, *81*(1), 65-92.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science*, *6*(1), 3-5.

Butterworth, B. (Ed.). (1983). *Development, Writing and Other Language Processes*. Academic Press.

Bybee, J. L. (1988). Morphology as lexical organization. In M. Hammond & M. Noonan (Eds.), *Theoretical morphology: Approaches in modern linguistics* (pp. 119–141). New York: Academic Press.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*(1), 55-67.

Domahs, U., Kehrein, W., Knaus, J., Wiese, R., & Schlesewsky, M. (2009). Event-related potentials reflecting the processing of phonological constraint violations. *Language and Speech*, *52*(4), 415-435.

Doron, E. (2003). Agency and voice: The semantics of the Semitic templates. *Natural Language Semantics*, *11*(1), 1-67.

Emmorey, K. (1989). Auditory morphological priming in the lexicon. *Language and Cognitive Processes, 4*(2), 73-92.

Ettinger, A., Linzen, T., & Marantz, A. (2014). The role of morphology in phoneme prediction: Evidence from MEG. *Brain and Language*, *129*, 14-23.

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research*, *1*(3), 183-192.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, 91-106.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, *22*(1), 179-228.

Frost, R., Forster, K. I., & Deutsch, A. (1997). What can we learn from the morphology of Hebrew? A masked-priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 829.

Frost, R., Deutsch, A., Gilboa, O., Tannenbaum, M., & Marslen-Wilson, W. (2000). Morphological priming: Dissociation of phonological, semantic, and morphological factors. *Memory & Cognition*, *28*(8), 1277-1288.

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22*(7), 615-621.

Graff, D. (2007). Arabic Gigaword Third Edition. Linguistic Data Consortium. Philadelphia, USA.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *Neuroimage*, *86*, 446-460.

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1-187.

Hale, J. (2001, June). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-8). Association for Computational Linguistics.

Halle, M. & Marantz, A. (1993). *Distributed Morphology and the pieces of inflection.* In: The View from Building 20, Ed. Hale, K. & Keyser, S. MIT Press, Cambridge, 111-176.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*(4), 434-446.

Janssen, N., Bi, Y., & Caramazza, A. (2008). A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language and Cognitive Processes*, *23*(7-8), 1191-1223.

Kielar, A., & Joanisse, M. F. (2011). The role of semantic and phonological factors in word recognition: An ERP cross-modal priming study of derivational morphology. *Neuropsychologia*, *49*(2), 161-177.

Koester, D., Holle, H., & Gunter, T. C. (2009). Electrophysiological evidence for incremental lexical-semantic integration in auditory compound comprehension. *Neuropsychologia*, *47*(8), 1854-1864.

Koester, D. (2014). Prosody in parsing morphologically complex words: Neurophysiological evidence. *Cognitive Neuropsychology*, *31*(1-2), 147-163.

Lewis, G., Solomyak, O., & Marantz, A. (2011). The neural basis of obligatory decomposition of suffixed words. *Brain and Language*, *118*(3), 118-127.

Marantz, A. (2013). No escape from morphemes in morphological processing. *Language and Cognitive Processes*, *28*(7), 905-916.

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, *101*(1), 3.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*,*10*(1), 29-63.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177-190.

McCarthy, J. J. (1983). Consonantal morphology in the Chaha verb. *Linguistics Department Faculty Publication Series*, 32.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006-1010.

Meunier, F., & Longtin, C. M. (2007). Morphological decomposition and semantic integration in word processing. *Journal of Memory and Language*,*56*(4), 457-471.

Mimouni, Z., Kehayia, E., & Jarema, G. (1998). The mental representation of singular and plural nouns in Algerian Arabic as revealed through auditory priming in agrammatic aphasic patients. *Brain and Language*, *61*(1), 63-87.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357.

Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*, *13*(1), 14-19.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., & Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland*.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*(6), 1090-1098.

Rastle, K., & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes*, *23*(7-8), 942-971.

Ryding, K. C. (2005). *A reference grammar of modern standard Arabic*. Cambridge University Press.

Salmelin, R. (2007). Clinical neurophysiology of language: the MEG approach. *Clinical Neurophysiology*, *118*(2), 237-254.

Sandra, D. (1994). The morphology of the mental lexicon: Internal word structure viewed from a psycholinguistic perspective. *Language and Cognitive Processes*, *9*(3), 227-269.

Schluter, K. (2013). Hearing words without structure: Subliminal speech priming and the organization of the Moroccan Arabic lexicon. Unpublished Dissertation. The University of Arizona.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*(1), 118-139.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400-2406.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, *26*(2), 100-107.

Seidenberg, M. S. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy?

Shaoul, C., Westbury, C. F., & Baayen, H. R. (2013). The subjective frequency of word n-grams. *Psihologija*, *46*(4), 497-537.

Solomyak, O., & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience*, *22*(9), 2042-2057.

Solomyak, O., & Marantz, A. (2009). Lexical access in early stages of visual word processing: A single-trial correlational MEG study of heteronym recognition. *Brain and language*, *108*(3), 191-196.

Stockall, L., & Marantz, A. (2006). A single route, full decomposition model of morphological complexity: MEG evidence. *The Mental Lexicon*, *1*(1), 85-123.

Twist, A. E. (2006). A psycholinguistic investigation of the verbal morphology of Maltese. Unpublished dissertation. The University of Arizona.

Uusvuori, J., Parviainen, T., Inkinen, M., & Salmelin, R. (2008). Spatiotemporal interaction between sound form and meaning during spoken word perception. *Cerebral Cortex*, *18*(2), 456-466.

Whiting, C., Shtyrov, Y., & Marslen-Wilson, W. (2014). Real-time functional architecture of visual word recognition. *Journal of Cognitive Neuroscience*

Wurm, L. H. (1997). Auditory processing of prefixed English words is both continuous and decompositional. *Journal of Memory and Language*, *37*(3), 438-461.

Figure captions:

Figure 1. Spectrogram of an example annotated stimulus item, [ضَجِرَ] *dajira - be angry*. The highlighted region corresponds to the position of critical consonant. Thin black lines indicate the phonetic boundaries that were assigned for the analysis.

Figure 2. Location of ROIs on inflated brain surface. A: Superior Temporal Gyrus. B: Transverse Temporal Gyrus. C: Grand average activity across all subjects and all words at 300 ms post third consonant onset; the scale represents the amplitude of activity above noise-level in dSPM units.

Figure 3. Incremental t-values across time when correlating morphological and linear measures of surprisal with neural activity in STG (above) and TTG (below), as time-locked to the onset of the third consonant. Critical t-values were computed on the full mixed-model as reported in the main analyses.

Figure 4. Average response time to the non-word conditions. Error bars represent 95% confidence interval.

Figure 5. Incremental t-values across time when correlating neural activity in the STG (above) and TTG (below) with the pair-wise differences between conditions using generalised linear hypothesis testing. Activity is time-locked to the onset of the second consonant.

Table captions:

Table 1. Average morphological and linear log. surprisal values.

Table 2. Correlation coefficients for experimental variables of interest. Morph = Morphological; Freq. = frequency; Surp. = Surprisal.