Lexical Networks in !Xung

Syed-Amad Hussain
Department of Computer Science
The Ohio State University
amadh881@gmail.com

Micha Elsner
Department of Linguistics
The Ohio State University
melsner0@gmail.com

Amanda Miller
[24]7.ai
San Jose, California
amanda.miller@247.ai

Abstract

We investigate the lexical network properties of the large phoneme inventory Southern African language Mangetti Dune !Xung as it compares to English and other commonly-studied languages. Lexical networks are graphs in which nodes (words) are linked to their minimal pairs; global properties of these networks are believed to mediate lexical access in the minds of speakers. We show that the network properties of !Xung are within the range found in previously-studied languages. By simulating data ("pseudolexicons") with varying levels of phonotactic structure, we find that the lexical network properties of !Xung diverge from previously-studied languages when fewer phonotactic constraints are retained. We conclude that lexical network properties are representative of an underlying cognitive structure which is necessary for efficient word retrieval and that the phonotactics of !Xung may be shaped by a selective pressure which preserves network properties within this cognitively useful range.

1 Introduction

We investigate the lexical network properties (LNPs) of the Southern African language Mangetti Dune !Xung (hereafter !Xung) as they compare to previously-studied languages. !Xung has 87 consonant phonemes, substantially larger than most of the world's languages (Miller, 2016; Miller-Ockhuizen, 2003; Dickens, 1994; Maddieson, 2013). Many of these sounds are clicks, typologically rare sounds found mostly in Southern Africa. In !Xung, close to 90% of content words begin with an initial click. While these properties place !Xung distinctly apart from most commonly-studied languages at the phonemic level, we analyze its lexical network (LN) to determine whether its mental lexicon is structurally different from those of languages with smaller inventories.

In a LN, as shown in Figure 1, nodes represent words and edges between nodes represent minimal pairs (Vitevitch, 2008). Vitevitch (2008) argues that the high connectivity and tendency toward clustering found in the English language lexicon are important aids to word learning and retrieval; later work finds similar properties in other lexicons (Arbesman et al., 2010; Shoemark et al., 2016). Some claims about the linguistic relevance of LNPs have been qualified by experiments showing that certain property values are inherent to the construction process of the network and can be replicated even when words are sampled from simple generative processes (Stella and Brede, 2015; Gruenenfelder and Pisoni, 2009; Turnbull and Peperkamp, 2016; Brown et al., 2018), though all these studies except Brown et al. point out that the LNs of natural languages maintain some distinctive properties.

Because !Xung has a very large phoneme inventory, it might in principle have very different network properties from previously studied languages. Any given word might have far more minimally different neighbors; alternately, the words might be spread out more thinly across a wider phonemic space. Our main questions in this study are (1) whether the network properties of !Xung differ from those of previously-studied languages, and, (2) if not, what phonological properties of the language lead to this network structure despite the large phoneme inventory?

Our initial analysis shows that most of the LNPs of !Xung lie within the range of values found for other languages in previous work. We next look at how these properties might vary over a range of lexicon sizes. Because large lexicons for !Xung are not available, we conduct these analyses on simulated data ("pseudolexicons") sampled from trigram models, following Gruenen-

Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 11–20

https://doi.org/10.18653/v1/P17

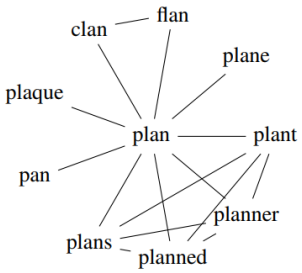

Figure 1: Example lexical network centered around the word "plan": (Turnbull and Peperkamp, 2016, Fig. 1).

felder and Pisoni (2009). Though this analysis must be considered preliminary due to the weakness of the trigram model, a comparison against the reported values from Shoemark et al. (2016) again finds no substantial difference. Having answered our first question, we turn to the second: we construct pseudolexicons with varying degrees of phonological structure, following Turnbull and Peperkamp (2016), and compare them to one another. We show that !Xung is more susceptible to the loss of phonotactic structure than English; simplistic sampling procedures create extremely unnatural lexicons due to the inventory size. To determine what phonotactic properties give the actual lexicon its shape, we create additional pseudolexicons that focus on specific phonological properties of !Xung. We find that pseudolexicons based on syllabic structure, including the syllable type inventory and co-occurrence restrictions on onset and rhyme within the syllable, move closer to the properties of the actual language, although a disparity still remains present. Overall, we find that !Xung has similar LNPs to previously studied languages. However, experiments with sampled lexicons show that when its syllable structure is disrupted, disparities between !Xung and English arise, hinting at a greater reliance on phonotactics to maintain the shape of the network.

2 Background

We conduct our analysis on LNs to derive cognitive and phonotactic conclusions. Vitevitch (2008) first presents this network model which assigns words as nodes and minimal pairs between these words as edges. He finds that lexical retrieval and language acquisition is aided by higher network density – largely defined by the network properties of assortative mixing and average clustering coefficient. Vitevitch (2008) and subsequent work on networks (Shoemark et al., 2016; Turnbull and Peperkamp, 2016) describe network structure in terms of four properties: **Fraction in Largest Island** is defined as the percent of the lexicon that is connected to the largest component, or island, in the network and characterizes the global connectivity of the network. The remaining three properties are calculated within this largest island: **Degree Assortativity Coefficient** shows the tendency of nodes to be connected to other nodes with similar degrees, where with higher values the central "hubs" of the network are connected to one another (Newman and Girvan, 2003); **Average Shortest Path Length (ASPL)** averages the minimum number of hops it takes to get between any two nodes in the largest island, similar to the game "Six Degrees to Kevin Bacon"; **average Clustering Coefficient (CC)** is defined as the number of edges that exist between neighbors divided by the number of possible edges between neighbors and can be thought of as "are my neighbors also neighbors with each other" or "do all my friends know each other?".

Later work on this model points out that network statistics are affected by lexicon size, phoneme inventory size, word length distribution, and the inclusion of morphological variants (Shoemark et al., 2016). Since these cannot all be controlled in cross-linguistic comparisons, indirect comparisons are often made. The phonological properties of the language can be used to generate pseudolexicons sampled from character language models, which are examined over several lexicon sizes. The trends for each language are then compared qualitatively against each other language.

Further work expands the use of pseudolexicons to determine the source of the network property statistics (Turnbull and Peperkamp, 2016). Instead of attempting to replicate the real phonotactic regularities of the language, pseudolexicons can vary in how many, and which, phonotactic properties of the original language they retain. By comparing several such pseudolexicons, Turnbull and Peperkamp (2016) conclude that the typical range of values of average CC are intrinsic to all LNs, typical values of largest island size and ASPL are determined by phonological rules, and degree assortativity may reflect some higher-level organization principle within the lexicon.

While this kind of previous research has established that some lexical network properties depend

on phonology, their experiments tell us relatively little about what specific phonological constraints have the greatest effect. In order to do so, we must move beyond comparing real languages to samples from generic statistical processes like ngram models and create distributions which enforce individual phonotactic constraints.

We employ a series of pseudolexicons which preserve various aspects of !Xung phonology to determine which phonological rules within these languages are responsible for preserving the typical values of largest island size and ASPL. We find that constraints on click placement and syllable structure can explain most, but not all the difference between randomly generated pseudolexicons and the real data.

3 Phonological Properties of !Xung

Mangetti Dune !Xung belongs to the Kxa language family (formerly known as the Northern Khoisan branch of the Khoisan family), and is a member of the Northern branch of the Juu subgroup, according to the classification of Sands (2003). The complete sound inventory of Mangetti Dune !Xung is provided in Miller (2016). Mangetti Dune !Xung contains 87 consonants, 45 of which are click consonants; its vowel inventory is also extremely large. There are only five contrastive vowel qualities, but there are many contrastive vocalic phonation types (modal, breathy, epiglottalized and glottalized), and the language also contrasts oral vs. nasal vowels. Nasality can combine with all the different phonation types, though there are some restrictions on which vowel qualities can combine with epiglottalization and nasalization. In addition, !Xung is a tone language; each mora may bear one of 4 distinct tone levels with some restrictions on their co-occurrence (Miller-Ockhuizen, 2003), leading to 7 possible contrastive tone patterns that occur on content words. (In our analysis, for purposes of determining minimal pairs, the tones are considered as contrastive features of the vowels.) Over 90% of content words in !Xung commence with a click consonant, while function words largely begin with a pulmonic (non-click) consonant.

Miller-Ockhuizen (2003) describes the phonology of a related Juu lect, Ju'hoansi. All native roots within Ju'hoansi (and !Xung) are either monosyllabic or bisyllabic with loan words constituting any trisyllabic roots. A syllable consists of an onset consonant followed by a 1 or 2-vowel nucleus with 2-vowel nuclei only occurring within the first syllable. The only coda consonants are nasals which end some monosyllabic roots. Within a word, 89 consonant types can occur in the initial position while only 4 types occur in medial position. Initial consonants are 91% pulmonic and velaric plosives, which includes all click types, with fricatives and nasal or liquid sonorants constituting the rest of the occurrences. Medial consonants are effectively limited to the sonorants β and ɾ (98% of medial consonants) and the nasals m and n. Guttural consonants and vowels only occur within the initial syllable and both never co-occur within the same syllable. The extensive co-occurrence restrictions in !Xung continue between tone and guttural vowels and consonants where, for instance, roots with partially epiglotallized vowels are always bitonal while roots with fully epiglottalized vowels are level toned. There are also several co-occurrence restrictions based on place of articulation with cross-height cross-place dipthongs only occurring in roots with back clicks and dipthongs with epiglottalized vowels and pharyngeal consonants causing the dipthongization of following front vowels. See Miller (2016) and Heikkinen (1986) for differences between Ju'hoansi and !Xung.

4 Basic properties

We begin by establishing the actual LNPs of the !Xung lexicon and comparing them to previous work.

4.1 Methodology

Our !Xung corpus contains 974 words, collected and transcribed into IPA as part of field work (Miller et al., 2008). For comparison, we use an English lexicon containing the 974 highest frequency words from the Fisher corpus (Cieri et al., 2004), converted to IPA using the CMU dictionary— though we do not believe that the !Xung lexicon contains strictly the most frequent words of the language, we do believe that the field workers chose to record words which they encountered frequently in storytelling and conversation.

We build each LN by assigning words as nodes and minimal pairs as edges. We build and analyze our networks using the python NetworkX package. From these networks, we derive the Fraction in Largest Island, Degree Assortativity, ASPL, and

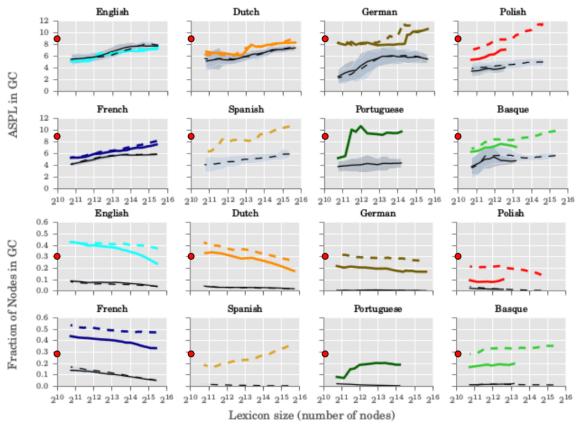

Figure 2: Two panels of Figure 4 from Shoemark et al., with superimposed dots for !Xung; colored lines show real language, dotted lines show pseudolexicons.

Property	!Xung	Closest value
% Lgst. Island	36.5	32 (Dutch)
ASPL	8.74	8 (German)
Deg. Assrt.	52.8	52 (German)
CC	52.4	35 (Polish)

Table 1: Lexical network properties of !Xung, along with closest comparison values from Shoemark et al. (2016).

Median degree	4
Mean degree	4.357
Min degree	1
Max degree	14
Degree std. dev.	2.749

Table 2: Degree (minimal pair) statistics of !Xung.

Average Clustering Coefficient. We then qualitatively compare these results to the values for the 8 reported languages[1] in Shoemark et al. (2016).

4.2 Results

The LN degree statistics are summarized in Table 2. Despite the potential for very large or small numbers of minimal pairs per word, we find that the actual maximum degree (number of pairs) is 14, the minimum is 1, and most words have about 4 neighbors. Lexical network properties are shown in Table 1. (Comparison values for the LNPs are derived from Shoemark's Figure 4 by reading the graph at the smallest lexicon size available; two panels of this graph are reproduced as Figure 2.) !Xung's value for Fraction in Largest Island falls within the observed range of variation; for Average Shortest Path and Degree Assortativity, !Xung represents an extreme of the observed values, but falls quite close to the measurements for German. Only for Clustering Coefficient is the !Xung value an outlier; !Xung is more tightly clustered than the other languages in the sample.

5 Analysis 1

The analysis above did not show definitive differences between !Xung and previously studied languages, but, as argued by Shoemark et al. (2016), LN measurements are best viewed as trends over several lexicon sizes rather than point measurements. With the limited data available for !Xung, we cannot obtain more than 974 actual words; instead, we follow previous work in using sampled data as a proxy. Though sampled data cannot be considered fully reliable, it can help us to understand whether !Xung phonology would probably create extreme LNP values if more data were available, or whether the outcomes would likely remain in the typical range.

5.1 Methodology

To create pseudolexicons that most accurately capture the phonotactics of each language, we use a trigram model with Ney's absolute discounting(Ney et al., 1994)[2]. Using these probabilities, we can extend the lexicon size by simulating "words" similar to those in the actual language.

We train the trigram models using the SRI Language Modeling (SRILM) Toolkit (Stolcke, 2002; Stolcke et al., 2011). We generate pseudolexicons of size 2^{10}, 2^{11}, 2^{12}, and 2^{13} for each language (we did not generate a 2^{13} length pseudolexicon for English) and average relevant network statistics over 50 trials.

5.2 Results

The trendlines appear in Figure 3. In an initial overview, we see that !Xung trend lines are similar to those for trigram-sampled English for most of the properties; Fraction in Largest Island trends upward (the network grows more connected), as does Degree Assortativity ("hubs" in the network

[1] Seven Indo-European languages of Europe: English, Dutch, German, Polish, French, Spanish and Portuguese—and one language isolate: Basque.

[2] Previous work used Kneser-Ney smoothing (Chen and Goodman, 1999); this is not as suitable for character-level modeling, since it uses a type-based backoff strategy designed for the sparsity of word rather than character statistics.

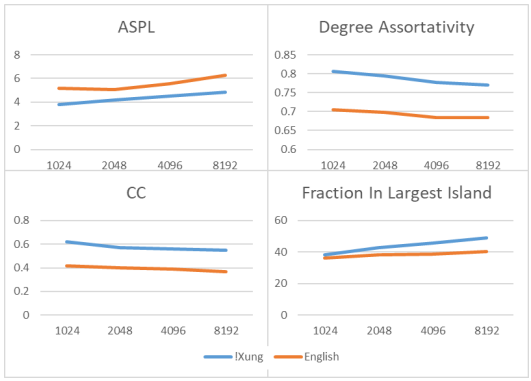

Figure 3: Trigram pseudolexicon network property values for trigram pseudolexicons of English and !Xung, and natural English data.

grow closer together), while clustering coefficient remains flat. The trend for ASPL differs (!Xung remains flat while English, like most languages in Shoemark's sample, increases). However, there are long flat intervals in Shoemark's trendlines for ASPL in German, Dutch and Portuguese.

For most of the LNPs, the slope of the natural English trendline is similar to that for trigram English, indicating that the language model is a reasonable proxy for additional data. However, for Fraction in Largest Island, the slope is reversed; real English grows less rather than more connected. This is probably due to the different lexical strata within English (less common words are often borrowings with different phonological patterns) (Shoemark et al., 2016). Without additional !Xung data, we cannot know how well the trigram LM corresponds to the real trendlines for !Xung, nor whether the real slope for Fraction in Largest Island would increase or decrease. However, increasing slopes are linguistically plausible; Spanish and Portuguese have increasing Largest Island sizes.

Overall, then, the trendlines for !Xung are plausibly within the range of variation shown by previously studied languages. The analysis below shows that trigrams are a poorer proxy for the !Xung lexicon than for English, generating a realistic size for the largest island but erroneous values for shortest path and assortativity, so these results must be taken with a substantial grain of salt. However, like the LNP statistics above, they represent converging evidence that !Xung's lexical network is not a linguistic outlier in the same way as its phonemic inventory.

6 Analysis 2

Analysis 1, like the preliminary examination, showed that the LNPs of !Xung broadly resemble those of previously studied languages. This raises the question: what phonological properties allow !Xung to have similar LNPs to these languages despite having a much larger phoneme inventory? In this analysis, we employ the methods used by Gruenenfelder and Pisoni (2009); Stella and Brede (2015); Turnbull and Peperkamp (2016) to create pseudolexicons with varying levels of phonological structure. A comparison of these pseudolexicons highlights the phonotactic disparities between !Xung and English.

6.1 Methodology

For each of our corpora, we generate the following pseudolexicons also used in Turnbull and Peperkamp (2016): Uniform – randomly selects from the phoneme inventory; Zipfian – randomly selects from the phoneme inventory given a Zipfian distribution; Scrambled – scrambles the phonemes of a word in place; Bigram – like the previously mentioned trigram LM; Trigram. We also create a Unigram pseudolexicon which randomly selects from the actual phoneme distribution. Pseudolexicons which sample single letters are given the same word length distribution as the original lexicon. Examples of words from these pseudolexicons are shown in Table 4 within the appendix. We compare the network properties of these pseudolexicons (averaged over 50 trials) within each language.

Since the pseudolexicons represent artificial distributions, which are known *a priori* to differ from the true distribution of words in the language, null hypothesis significance testing is inappropriate to assess the degree of difference— an arbitrarily small p-value could always be obtained by sampling more data. Instead, we use Cohen's d as a measure of the effect size; d measures the difference between means scaled by the standard deviation.

6.2 Results

Table 3 shows the results, also plotted in Figure 4. The lower right panel shows that only the bigram and trigram lexicons generate realistic sizes for the largest island. Other pseudolexicons are highly disconnected. This is especially the case for !Xung relative to English; for instance,

	Unif	Zipf	Scram	Unig	Bigr	Trigr	Natural
% Lgst. (mean)	0.7	24.9	6.1	9.0	41.7	37.7	36.6
% Lgst. (d)	-246.9	-6.0	-15.1	-15.3	2.6	0.6	
% Lgst. (mean)	8.6	32.1	12.5	18.3	33.3	36.6	38.4
% Lgst. (d)	-30.3	-4.6	-13.4	-9.8	-3.5	-1.1	
ASPL (mean)	2.0	4.4	5.8	5.9	3.9	3.7	8.7
ASPL (d)	-17.9	-13.6	-1.9	-3.1	-24.0	-23.7	
ASPL (mean)	5.1	4.2	6.0	5.9	4.6	5.2	6.1
ASPL (d)	-1.4	-12.1	-0.1	-0.4	-7.5	-3.1	
DA (mean)	-27.0	35.1	26.9	31.5	80.1	80.5	52.8
DA (d)	-3.7	-2.6	-2.1	-2.0	11.2	14.4	
DA (mean)	39.1	32.6	46.8	45.8	71.1	72.3	43.6
DA (d)	-0.5	-3.6	0.5	0.5	7.8	8.0	
CC (mean)	35.6	58.7	50.1	48.9	58.6	60.8	52.4
CC (d)	-0.7	2.1	-0.4	-0.7	3.5	4.2	
CC (mean)	37.3	44.1	39.0	36.3	42.9	42.5	36.5
CC (d)	0.2	4.1	0.7	-0.1	3.5	2.9	

Table 3: Pseudolexicon LNPs (!Xung in white, English in gray); mean and Cohen's d versus the natural language.

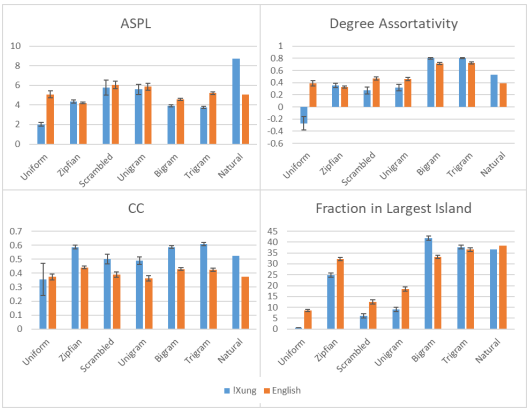

Figure 4: Network property values for pseudolexicon models of English and !Xung, ordered by phonotactic similarity to the natural language, with right-most being the natural language itself.

the English uniform pseudolexicon (far left) has nearly 10% of the nodes in the largest island, while !Xung has essentially none on average, with extremely high variance. This disparity between languages is caused by the large phonemic inventory, which creates fewer minimal pair matches when randomly sorted, as in the uniform, Zipfian, scrambled, and unigram pseudolexicons. !Xung thus serves as a counterexample to recent claims that simplistic random lexicons created with unigram sampling can mimic the properties of real LNs (Brown et al., 2018). The disparity begins to shrink as the pseudolexicons become more natural, suggesting that disparities due to the large phonemic inventory are reduced by phonological structure and that phonotactic constraints on word forms in !Xung lead the lexicon to include more minimal pairs.

For the lexicons with reasonable island sizes, the values of clustering coefficient are relatively stable across all pseudolexicons, as in Turnbull and Peperkamp (2016), though again highly variable for the Uniform lexicon. The shortest path and associativity measures show that Bigram and Trigram lexicons are more compact and centralized than the actual lexicon for both !Xung and English (paths are shorter and associativity is higher). However, these differences are larger for !Xung than for English as measured by Cohen's d. The Unigram and Scrambled lexicons, meanwhile, are more realistically dispersed, but also disconnected ($<$ 10% of nodes in the largest island).

Overall, the results show that the network structure of !Xung reflects properties which go beyond the frequencies of individual segments, including some characteristics which are poorly captured even by trigram models. The differences between simple pseudolexicons and the real properties of the language are generally greater (in terms of d) for !Xung than for English. Thus, we conclude that the !Xung network is less resilient to phonotactic disruption than the English network.

	English	!Xung
Uniform	ŋsθ, iɛhh, iʒgmʊ	ü‡^{h}è
Zipfian	oʊpə, ɔəɛŋə, ŋɔæɔs	óm̀n|n̊|^{h}, g‖ŋ̀!, g|ho
Scrambled	ɛθlih, əwrdnəɪŋ, krljəɛaən	ò‖ŋ, eŋ̊!^{h}e, äà^{h}!
Semi-Scrambled	(N/A)	‖ŋò, ŋ̊!^{h}ee, !à^{h}ä
CV	ɛnfəzamɛs, vɪlɔk-itdr, nnaəln	nő!'ùŋ|, g‖áχúβ, ʃínòʒ
KCV	(N/A)	ŋ!dóm, ǂ'zùs', g‖lós
Free O+R	(N/A)	!ònä, dʒä, g‖ö^{ʔn}
Pos. O+R	(N/A)	gǂó, ŋ!ò^{ʔ}và, wì
Syllable	(N/A)	n̊ŋ^{h}úí^{n}, ‖àò^{n}, g!ə̀mfiè
Unigram	wjork, sɛgit, njŋe	úŋ́‖úlùb, téùχà^{ʔ}à, íl‖!ìò
Bigram	rnɔsm, ɔjə, tekjks	!?ŕń, má^{ʔ}ìì, |'ùβä^{n}
Trigram	plæŋ, lɪŋ, hæv	n̊ŋ^{h}ùì, n̊ŋ^{h}ülà, tä^{h}
Natural Lexicon	hɛlθi, wəndərɪŋ, kɛrəlajnə	‖èŋ, ŋ̊!^{h}ee, !ää^{h}

Table 4: Three random words from pseudolexicons.

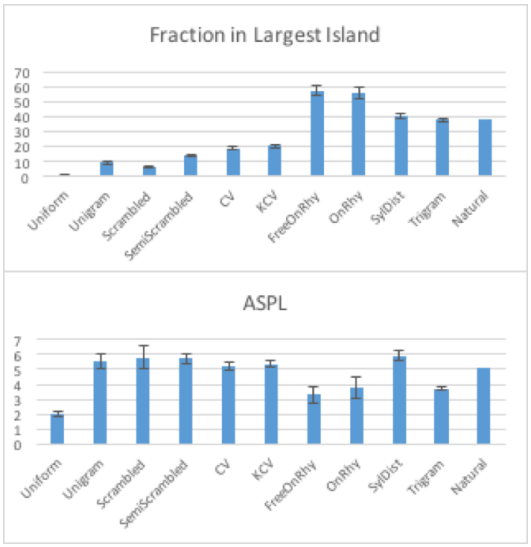

Figure 5: LNPs in Analysis 3.

7 Analysis 3

We continue our investigation by attempting to determine which phonotactic properties of !Xung might be most important in maintaining its structure. Several properties of !Xung phonology might be important in constraining its network structure. These include its relatively simple syllabic structure, the positional constraints on initial medial consonants, and the co-occurrence restrictions on consonants and vowels within the syllable. To highlight these properties, we compare our Scrambled pseudolexicon to pseudolexicons designed to respect some of these properties.

7.1 Methodology

In the CV lexicon, we extract a distribution over word templates by transforming each consonant into C and each vowel into V, then generate each word by sampling a template from this distribution and filling it with random consonants and vowels sampled from the unigram distribution. The CV pseudolexicon forces the generated words to contain reasonable proportions of vowels and consonants, but it does not enforce any positional constraints; words may contain unnatural features like illegal codas, sequences of vowels which do not form diphthongs, and medial clicks. We next test the effect of the constraint that !Xung content words tend to begin with a click, by generating a Semi-scrambled lexicon (scrambling each real word in place, but any present click stays at the initial position) and KCV (like CV, except that the initial syllable will begin with a click and subsequent ones will only contain vowels and pulmonic consonants). Examples of words from these pseudolexicons can be found in Table 4.

We next syllabify the !Xung corpus (treating each sequence of vowels as a syllable nucleus and maximizing onsets). We use this database of syllables to create a sequence of pseudolexicons which sample larger prosodic units rather than single segments. These pseudolexicons are length-matched to the original corpus in number of syllables (not segments). Free Onset+Rhyme forms syllables by sampling attested onsets and rhymes, but draws them from anywhere in the corpus, ignoring positional constraints. Positional Onset+Rhyme enforces placement constraints on consonants by sampling word-initial, medial and final onsets and rhymes from separate distributions. Finally, Syllable samples whole syllables from the correct positional distributions, enforcing the co-occurrence constraints between onsets and rhymes as well as the placement constraints.

7.2 Results

The results (averaged over 50 trials) are shown in Table 5. The CV lexicon, which forces words to contain realistic proportions of vowels and consonants, roughly triples the largest island size versus the Scrambled lexicon, but does not create a realistic LN. Forcing clicks to occur only at word beginnings does relatively little; neither the Semiscrambled nor KCV lexicons look very different from the CV lexicon. Lexicons formed using ac-

	Scram	Semi	CV	KCV	Free Ons+Rhy	Posit. Ons+Rhy	Syll	Trigr	Natural
% Lgst. (mean)	6.1	13.6	18.6	20.3	57.4	55.7	40.6	37.7	36.6
% Lgst. (d)	-15.1	-17.2	-9.2	-8.4	3.5	2.6	1.2	0.6	
ASPL (mean)	5.8	5.7	5.2	5.4	3.3	3.8	5.9	3.7	8.7
ASPL (d)	-1.9	-5.0	-5.5	-7.6	-5.1	-3.6	-4.1	-23.7	
DA (mean)	26.9	38.8	43.8	43.0	47.0	43.6	59.6	80.5	52.8
DA (d)	-2.1	-3.7	-1.1	-1.3	-1.1	-1.1	2.5	14.4	
CC (mean)	50.1	55.2	47.4	46.1	63.1	62.9	51.2	60.8	52.4
CC (d)	0.0	0.0	-0.1	-0.1	0.1	0.1	-0.1	0.1	

Table 5: Statistics of phonotactically targeted pseudolexicons (mean and Cohen's d versus natural language).

tual onsets and rhymes attested from the corpus have much larger island sizes— in fact, larger than the real graph (56% vs 36%). The Positional Onset+Rhyme model is quite similar to the Onset+Rhyme model across all the LNPs. Finally, the Syllable lexicon has a realistic island size (41%), and is wider than the Onset+Rhyme or Trigram networks, with an average path length of 5.9 (vs 3.3-3.8, compared with the actual 8.7)[3].

Comparing the Onset+Rhyme models to the CV lexicon, we find that the syllabic structure of !Xung helps to ensure that the network is connected. Surprisingly, positional constraints on consonant placement (clicks at the beginning, restricted set of medials) have a limited impact on the shape of the network; KCV is similar to CV, and Positional Onset+Rhyme to Free Onset+Rhyme. However, the co-occurrence restrictions on onset and rhyme within the syllable, for instance, constraints on gutturals, are important in limiting connectedness and creating the long shortest-path distances of the real lexicon, since both these properties appear only in the pseudolexicon which samples whole syllables as units. Co-occurrence restrictions within the syllable widen the LN by preventing the formation of a minimal pair which would be phonologically unnatural, since its onset and rhyme would not match.

8 Conclusion

Overall, we find that the network properties of !Xung do not substantially differ from previously studied languages despite fundamental phonological disparities. This supports the argument of Vitevitch (2008) that the LNPs indicate an underlying cognitive structure. Vitevich proposed that the global shape of the network enables efficient word learning and retrieval from memory; it is also plausible that the network structure is necessary to avoid confusing large numbers of minimal pairs in auditory perception. In any case, the preservation of this global structure suggests a selective pressure shaping the phonotactics of these languages (and others with large inventories) — phonotactic rules may arise and change over time in ways that preserve the network properties within a cognitively useful range. For instance, the differences between randomly scrambled and syllabic pseudowords indicate that the restricted syllable inventories of !Xung and Ju may force words to cluster more tightly in the LN, compensating for the large number of contrastive phonemes. In other words, the underlying universal structure may be, not linguistic, but cognitive. This universal architecture may require certain patterns of connectivity within the lexicon, and these, in turn, may entail particular phonotactic patterns.

Looking forward, we plan to expand our current LN analysis to include data from relatives of !Xung such as Ju|'hoansi,[4] as well as languages with small phoneme inventories, such as certain Polynesian languages[5]. Through this, we hope to uncover how our hypothesis operates across a range of inventory sizes and types.

Additionally, we plan to investigate the func-

[3]Cohen's d estimates a slightly *larger* effect separating the ASPL for Syllable from the natural language than Onset+Rhyme; this is because Syllable has a substantially lower variance.

[4]We conducted a preliminary analysis of a 3733-word lexicon of Ju|'hoansi collected by Biesele et al. (2006) and found similar results to those we obtained from !Xung. However, the IPA transcription of this data is not consistent, so we have chosen not to present it here.

[5]Hawai'ian was previously studied by Arbesman et al. (2010) who found a comparatively larger giant component and shorter average path lengths than several other languages; however, they did not control for lexicon size.

tional load and potential confusability of !Xung phonemic contrasts. A LN assumes real world speakers can distinguish perfectly between minimal pairs. However, with the large phoneme inventory of !Xung, these clicks may be confusable in real speech, cf. (Fulop et al., 2004). We hope to determine how the network properties change when potential confusions between sounds are taken into account.

Acknowledgements

We thank Rory Turnbull, Philippa Shoemark, Eric Fosler-Lussier, the attendees of OSU's Workshop on the Emergence of Linguistic Universals and Phonies discussion group, and six anonymous reviewers for their many helpful comments and suggestions. This work was funded by NSF 1422987 to the second author.

References

Samuel Arbesman, Steven H Strogatz, and Michael S Vitevitch. 2010. The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, 20(03):679–685.

M. Biesele, B. C. Boo, H. K. Gcao, G≠kao M. /K, Kaqece K N!., A. Miller, /A. F /Kunta, and C. N! Tsamkxao F. /U. /Ui. 2006. *Ju|'hoansi Dictionary, Revised version of Dickens, P. Ju|'hoan-English – English-Ju|'hoan Dictionary*. unpublished manuscript, The Kalahari People's Foundation and The Ju—hoan Transcription Group.

Kevin S. Brown, Paul D. Allopenna, William R. Hunt, Rachael Steiner, Elliot Saltzman, Ken McRae, and James S. Magnuson. 2018. Universal features in phonological neighbor networks. *Computing Research Repository*, arXiv:1804.05766. Version 1.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of LREC*, volume 4, pages 69–71.

P. Dickens. 1994. *Ju|'hoan-English English-Ju|'hoan Dictionary*. Koppe.

Sean A Fulop, Peter Ladefoged, Fang Liu, and Rainer Vossen. 2004. Yeyi clicks: Acoustic description and analysis. *Phonetica*, 60(4):231–260.

Thomas M Gruenenfelder and David B Pisoni. 2009. The lexical restructuring hypothesis and graph theoretic analyses of networks based on random lexicons. *Journal of Speech, Language, and Hearing Research*, 52(3):596–609.

Terttu Heikkinen. 1986. Outline of the phonology of the !Xũ dialect spoken in Ovamboland and western Kavango. *South African journal of African languages*, 6:18–28.

Ian Maddieson. 2013. Vowel quality inventories. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology.

A. Miller, L. Namaseb, S. Sands, S. Shah, M. Aromo, C. Augumes, R. Fransisko, T. Kaleyi, D. Prata, and S. Riem. 2008. *Mangetti Dune !Xung Dictionary*. unpublished manuscript, The Ju|'hoan Transcription Group, The Kalahari People's Foundation, The University of Namibia and Northern Arizona University.

Amanda L Miller. 2016. Posterior lingual gestures and tongue shape in Mangetti Dune !Xung clicks. *Journal of Phonetics*, 55:119–148.

Amanda Miller-Ockhuizen. 2003. *The Phonetics and Phonology of Gutturals: A Case Study from Ju|'hoansi*. Outstanding Dissertations in Linguistics Series. Routledge.

Mark EJ Newman and Michelle Girvan. 2003. Mixing patterns and community structure in networks. In *Statistical mechanics of complex networks*, pages 66–87. Springer.

Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1–38.

Bonny Sands. 2003. Juu subgroups based on phonological patterns. In *Khoisan languages and linguistics: Proceedings of the 1st International Symposium*, pages 85–114.

Philippa Shoemark, Sharon Goldwater, James Kirby, and Rik Sarkar. 2016. Towards robust cross-linguistic comparisons of phonological networks. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 110–120.

Massimo Stella and Markus Brede. 2015. Patterns in the english language: phonological networks, percolation and assembly models. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(5):P05006.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proceedings of IEEE automatic speech recognition and understanding workshop*, volume 5.

Rory Turnbull and Sharon Peperkamp. 2016. What governs a language's lexicon? Determining the organizing principles of phonological neighbourhood networks. In *International Workshop on Complex Networks and their Applications*, pages 83–94. Springer.

Michael S Vitevitch. 2008. What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, 51(2):408–422.