Statistical learning based speech segmentation: A cross-linguistic, corpus-based perspective

Michael Fry
Univeristy of British Columbia

Abstract: This paper investigates the statistical separability of within-word and between-word segment transitions in spontaneous speech corpora. Three metrics previously proposed in the literature, Forward Transitional Probability, Backward Transitional Probability and Mutual Information are employed to encapsulate the statistical regularities in each corpus that are thought to enable the separation of these transition types. It has been claimed that infants use such statistical information, available to them via a statistical learning mechanism, to segment continuous speech into words during the first stages of language acquisition. Four corpora are analyzed with results providing evidence that statistical separability of within-word and between-word transitions does exist, to varying degrees, cross-linguistically. Further, while no one metric consistently affords the most separability, Mutual Information is generally the most robust.

Keywords: Statistical Learning, Corpus Linguistics

1 Introduction

Segmenting continuous speech into discrete, meaningful chunks is foundational for acquiring language. However, with no consistent acoustic cues demarcating words in natural speech (Cole and Jakimik 1980; Lehiste 1960), the mechanisms underpinning word segmentation in infants are not fully known. Statistical learning, as developed in Saffran, Aslin and Newport's (1996a) seminal work, is one possible mechanism which relies on an infant's ability to extract statistical regularities from stimuli. Speech segmentation based on statistical learning has been seen in infants as young as 5.5 months (Johnson and Tyler 2010) and has been shown to precede stress patterning, another known segmentation strategy (Thiessen and Saffran 2003). These findings, however, arise solely from artificial language learning experiments, which in turn has lead some researchers to question the effectiveness of statistical learning with more natural language (Johnson and Jusczyk 2001; Johnson and Tyler 2010; Yang 2004). One avenue to address this question is to investigate the statistical regularities present in natural speech and quantify how useful such information actually is for the purpose of speech segmentation. To this end, this paper reports on the statistical regularities in four natural language speech corpora and evaluates their utility for segmenting speech.

To encapsulate the statistical regularities in speech, Saffran et al. (1996a) employ *Transitional Probability* – a measure of the likelihood of unit y following unit x in a corpus (speech or otherwise). The authors argue that the tracking of Transitional Probabilities (TPs) allows infants to delineate words in continuous speech. They state that TPs are, in general, higher for within-word transitions than between-word transitions, as words are made of units (e.g. segments, syllables) that regularly occur together. Transitional Probability has been used extensively in the speech segmentation literature (see de la Cruz Pavía (2012) for an overview), has been applied to both segment and syllable transitions (Daland 2009; Toro et al. 2005), and has been used to identify word-boundaries in transcribed corpora with limited success (e.g. Swingley 1999).

Contact info: mdfry20@gmail.com

In UBC Qualifying Papers 3 (2013–2015),
University of British Columbia Working Papers in Linguistics,
Andrei Anghelescu, Joel Dunham, and Natalie Weber (eds.), 2016.

A related alternative to TP is the metric of Mutual Information, which calculates the shared information of adjacent units. Mutual Information (MI) was originally applied to speech segmentation by Brent (1999a) and is increasingly being used alongside, or in place of, TP in the literature (Brent 1999b; Daland 2009; Rytting 2004; Swingley 2005). This trend is the result of studies such as Swingley (1999) and Rytting (2004) that showed MI, as compared to TP, allows for more accurate word-boundary identification with fewer false word-boundary predictions in English and Modern Greek corpora, respectively.

The current work uses TP and MI to compare within-word and between-word segment transitions in phonemically-transcribed spontaneous speech. Crucially, this paper reports on the *separability* of the two transition types (within-word, between-word) based on statistical regularities as encapsulated by these metrics; the work does not consider how accurately the metrics predict word-boundaries in speech corpora. This was decided as a model of speech segmentation is not necessary to address the question of the utility of such statistical regularities, separability is sufficient. The separability of transition types is vital for word segmentation as, if there were no separability, words could not be delineated by statistical learning. The statistical regularities used are meant to be analogues of regularities arrived at by an infant's statistical learning mechanism. Finally, as statistical learning based speech segmentation is not meant to be language specific (Saffran et al. 1996a), within-word and between-word transitions are compared in four languages. This is done to enhance the generalizability of results, to enable inquiry into cross-linguistic patterns, and to provide a starting point for identifying what ranges of metric values actually exist in natural language.

Details of the corpora and explicit metric definitions with example calculations are provided first. Next, the methodology of how each corpus was processed and how separability of transition types was analyzed is reported. Finally, results are presented and their consequences for the statistical learning based speech segmentation are discussed.

2 Corpora and language choices

Four corpora are used in this study: The Buckeye Corpus of Conversational Speech (BCCS) (Pitt et al. 2007), the Corpus of Spontaneous Japanese (CSJ) (Maekawa 2003), the Hong Kong Cantonese Corpus (HKCC) (Leung and Law 2001) and the Tunisian Arabic Railway Interaction Corpus (TARIC) (Masmoudi et al. 2014). These corpora contain spontaneous speech from Ohioan English, Tokyoite Japanese, Hong Kong Cantonese and Tunisian Arabic, respectively. Spontaneous speech corpora were chosen because the current project is concerned with statistical regularities in natural speech. Adult speech was chosen over infant-direct speech because infants are likely able to extract regularities from speech in their surroundings, even if it is not explicitly directed towards them. Support for this comes from the observation that attention attentuates statistical learning (Toro et al. 2005) and infants are capable of attending to noise in their environment.

As statistical learning based speech segmentation is thought to occur cross-linguistically (Saffran et al. 1996a) and prior to other segmentation strategies (Mattys et al. 2005; Thiessen and Saffran 2003), statistical separability of within-word and between-word transitions is expected across languages. To scrutinize this, separability is considered in four languages. Languages were selected to enhance the generalizability of results; that is, they were chosen to ensure variability in linguistic dimensions that likely affect the regularity of segmental patterning. Table 1 provides a summary of this variability compiled from language phonologies (Bauer and Benedict 1997; Gibson 1998; Ito and Mester 1995; Roach 2009).

Table 1: A summary of the variability between languages

	English	**Cantonese**	**Japanese**	**Tunisian Arabic**
Morphological Typology	Fusional/Analytic	Analytic	Agglutinative	Fusional
Syllable Maximum	CCCVVCCCC	CVVC	CV{V}{C}	CCVCCC
Tone/Accent	Lexical Stress	Lexical Tone	Pitch-Accent	Lexical Stress
Phonemic Inventory Composition	Cs $\approx$ 25 Vs $\approx$ 18	Cs $\approx$ 22 Vs $\approx$ 13	Cs $\approx$ 17 Vs $\approx$ 5	Cs $\approx$ 36 $\approx$ 3
Language Family	Indo-European	Sino-Tibetan	Japonic	Semitic

These dimensions were chosen with the following motivations. Morphologically rich languages (e.g. Japanese) contain many between-morpheme boundaries that may be confusable with between-word boundaries (as they occur frequently in many different contexts). Restricted syllable phonotatics are correlated with more syllables-per-word (Pellegrino et al. 2011), which may result in lower metric values for within-word transitions and make them more similar to between-word transitions. Tone/Accent allows for segmental homophones to be semantically distinct, which may increase the number of within-word segment transitions, resulting in higher metric values for within-word transitions and making them more separable from between-word transitions. Finally, phonemic inventory composition affects the number and type of transitions possible (e.g. fewer vowels means an increase in the frequency of transitions containing those vowels), which may result in large variability of metric values for transition types. If between-word and within-word transitions are separable cross-linguistically in spite of this variablity, it is reasonable to infer such separability serves an important function in language.

Other dimensions of language variability likely affect the separability of transition types and may even better predict language patterning. However, these dimensions provide a reasonable starting point for this preliminary investigation. In future work, more specificity for morphological system and syllable shape (perhaps average syllable size) may provide better descriptors which tie into the current analyses.

2.1 Predictions

Considering the variability in Table 1, the following patterns are predicted: Japanese, as a morphologically rich language, will have less separability of transition types than other languages; Cantonese, as a tone language, will have more separability of transition types than other languages; and, English and Cantonese will pattern together in some way and Japanese and Tunisian Arabic will pattern together in some way given their more similar phonemic inventory compositions.

3 Metrics

Two instantiations of Transitional Probability (Forward and Backward) and one of Mutual Information are used herein. Forward Transitional Probability (FTP) is a measure of the likelihood that a unit y will follow a given preceding unit x. Backward Transitional Probability (BTP) is a measure of the likelihood that a unit x will precede a given following unit y. It is important to note

that while the direction of transition differs, the function of the metrics for the purpose of speech segmentation is the same. Recall that it is the *tracking* of TPs which is thought to enable speech segmentation – between-word transitions, in general, should have lower TPs than within-word TPs (Saffran et al. 1996a) – and FTPs and BTPs are similarly trackable. In this paper, the units used are either monophones or diphones.

Unlike TP, Mutual Information is not a measure of the probability of one unit transitioning to/from another; it is a measure of the shared information between units in an Information-Theoretic sense and is often reported as bits of information (Swingley 2005). A detailed understanding of Information Theory (Shannon 1948) is not necessary here; for our purposes, MI is interpretable as a measure of the dependency two units have on each other (i.e. how frequently they occur together). In this way, MI should, in general, be higher for within-word transitions than between-word transitions. This parallel to TP entails that MI values can be tracked in much the same way. Formal definitions of the three metrics are provided in Figure 1.

$\text{FTP}(xy) = \text{FTP}(y\|x) = \frac{p(xy)}{p(x\_)}$	$\text{BTP}(xy) = \text{BTP}(x\|y) = \frac{p(xy)}{p(y\_)}$	$\text{MI}(xy) = \log_2 \frac{p(xy)}{p(x\_)p(y\_)}$
(Saffran et al. 1996b)	(Perruchet and Desaulty 2008)	(Brent 1999a)

Figure 1: Formulas for FTP, BTP and MI

It is worthwhile to compare these metrics as they are all commonly used in the literature and there is an on-going discussion of which is most effective. While there is a contrast in numerical units (probability and bits), the separability of within-word and between-word transitions that each metric affords is quantifiable and therefore directly comparable. As MI incorporates parts of both FTP and BTP, it will naturally share in the successes and failures of each of them. This may seem to make MI redundant, however the interaction actually provides MI with a unique behaviour that is not captured by simply considering FTP and BTP.

3.1 Calculation examples

For each metric, x and y can be any unit and thus are definable by the user. If we consider the sentence in (1) as a speech corpus, one could calculate the BTP for the transition between the segments /ð/ and /ə/ in the following way.

(1) /ð.ə.k.w.ɪ.k.b.ɹ.aʊ.n.f.ɑ.k.s.dʒ.ʌ.m.p.t.oʊ.v.ə.ɹ.ð.ə.l.eɪ.z.i.d.ɑ.g/

Setting x to be /ð/ and y to be /ə/, then:

$$\text{BTP}(/\text{ðə}/) = \frac{p(/\text{ðə}/)}{p(/\_\text{ə}/)}$$

Here, $p(/\text{ðə}/)$ represents the probability (relative frequency) that any two adjacent segments (i.e. any diphone) are /ðə/ and $p(/\_\text{ə}/)$ represents the probability that any given diphone ends with /ə/ in the corpus. There are 31 total diphones, of which 2 are /ðə/, and 3 end with /ə/. Therefore:

$$\text{BTP}= \frac{p(/\text{ðə}/)}{p(/\_\text{ə}/)} = \frac{(\frac{2}{31})}{(\frac{3}{31})} \approx 0.667.$$

Similarly, one could calculate MI(/ðə/) as:

$$\text{MI}(/\text{ðə}/)= log_2 \frac{p(/\text{ðə}/)}{p(/\text{ð}\_/)p(/\_\text{ə}/)} = log_2 \frac{(\frac{2}{31})}{(\frac{2}{31})\frac{3}{31}} \approx 3.37$$

An important characteristic of these metrics is that they rely solely on the relative frequencies of the *units* in the corpus. Since **frequency exists independently of location**, every repeated transition (e.g. /ðə/ in (1)) has one value per metric. Further, the metric values remain the same regardless of whether the transition occurs only within-word, only between-words or both. While this fact necessarily leads to the confusability of whether a transition demarcates words or not, there is often one transition type which occurs much more frequently than the other[1], enabling correct segmentation more often than incorrect segmentation (Cairns et al. 1997).

4 Methodology

4.1 Pre-processing corpora

To ensure the same method for calculating FTP, BTP and MI could be used with all languages, the format of all corpora was unified. To do this, each corpus was imported into the recently available Phonological CorpusTools (PCT) (Hall et al. 2015) using its *Import Spontaneous Speech Corpus* function. The function has several options to handle importing a variety of data files: the TARIC and HKCC corpora were imported as running text files (*.txt*), the CSJ was imported from Praat TextGrids (Boersma and Weenink 2014) and the BCCS was imported using PCT's default *Import BCCS* option. After import, PCT creates a standardized *.corpus* file which can be interacted with either through PCT's graphical interface, or through command-line functions. This unified format includes indices corresponding to discourses, speakers, sentences, words, transcriptions along with a variety of others. The format was necessary to ensure the same Python script could be used for each corpus. *Words* were taken as-are from each corpus as delimited by whitespace; the whitespace corresponds to breaks between orthographic words.

4.2 Extracting transitions

Diphone and triphone transitions were extracted from each corpus. Diphones were chosen following the lead of Daland (2009) who argues for their significance in word segmentation (e.g. /pd/ virtually never occurs within-word and thus is a good indicator of a between-word transition). Triphones were chosen following common practice in speech processing and recognition (e.g. Glass 2003) and necessarily require two variations because between-word transitions could be either between the first two segments and the third (2↔1) or the first segment and the following two (1↔2). This distinction directly affects metric calculations as x and y correspond to monophone or diphone units, relativized to their occurrences in triphones, depending on condition. For example, in transition condition triphone[2↔1], $p(x)$ would be the relative frequency of triphones beginning with diphone

[1]Consider the English word *the*, where /ðə/ occurs within-word, and the sentence *I loathe umbrellas*, where /ðə/ occurs between-word. Nonetheless, *the* is much more common, and overall segmentation accuracy would remain quite high if /ðə/ was consistently considered a within-word transition.

x, and $p(y)$ would be the relative frequency of triphones ending with monophone y. This mismatch of diphones/monophones results in a wide range of probability values.

To be explicit, it is beneficial to highlight the distinction between *transition condition* and *transition type*. Transition condition refers to the conditions diphone[1↔1], triphone[2↔1], and triphone[1↔2] as described here. Transition type refers to the distinction between within-word and between-word transitions. Thus, the separability of transition types is to be analyzed for each of the three transition conditions.

A Python script was written to extract all diphone and triphone transitions which occurred in each corpus. Table 2 is an example of a subset of the set of diphone transitions from the CSJ.

Table 2: Segment transitions, their frequency of occurrence between-words (*bw*) and within-words (*ww*), and their corresponding values for each metric. Data are from the CSJ.

1	transition	raw_freq_bw	raw_freq_ww	ftp	btp	mi
2	e.r	1039.0	6155.0	0.07277398992453517	0.13850330182322249	2.05220101604292
3	s.eH	0.0	3500.0	0.07250429846912353	0.23420770877944325	7.106420706979233
4	iH.o	115.0	0.0	0.029963522668056276	0.0007282998315410824	0.27794525082209254
5	i.o	2591.0	381.0	0.02738665683744932	0.01882180086382693	0.25404193252545937
6	o.eH	1001.0	25.0	0.006497454213846036	0.06865631691648821	0.6368400790414975
7	kj.i	0.0	8876.0	0.6943053817271589	0.08179212856734766	9.371276828137384
8	o.by	38.0	0.0	0.00020464645236466804	0.1784037558685446	1.6548318798810882
9	c.uH	0.0	355.0	0.03407236778961512	0.022420108627005178	3.151851105473632
10	oH.e	125.0	247.0	0.011089581159636309	0.003763163485174957	0.1643157935214933
11	u.d	1900.0	543.0	0.04070309896701099	0.05766008166348037	1.407126949249744

Segments are delimited in PCT by periods, thus all diphone transitions are of the type $x.y$. Frequency is split into within-word (ww) and between-word (bw) occurrences for later processing (Section 4.3), but, as mentioned previously, all metrics rely on total frequency of the units in the corpus (i.e. the sum of within-word and between-word raw frequencies). Each metric value was calculated for every transition using the formulas described in Section 3 and added into their respective rows.

4.3 Isolating transition types

From the data in Table 2, the **distributions** of each metric value for each transition type was constructed using the raw frequencies of occurrence. If a transition occurred both within-word and between-words, it contributed to the construction of both transition type distributions. Frequency was incorporated in the distributions by reduplicating metric values as many times as they occurred for each type. As an example, the transition $[e.r]$ contributed 6155 corresponding FTP, BTP and MI values to within-word distributions, but only 1039 of each value to between-word distributions (for Japanese diphones). Similarly, $[iH.o]$ contributed only to between-word distributions and $[kj.i]$ contributed only to within-word distributions. The result of this process was the creation of **metric value distributions** for each transition type, for each metric.

4.4 Summary of methodology

Four corpora (BCCS, CSJ, HKCC, TARIC) were processed for segment transitions in three transition conditions (diphone[1↔1], triphone[2↔1], triphone[1↔2]). Three metric values (FTP, BTP,

MI) were then calculated for all transitions in each of the transition conditions. Distributions of metric values for each transition type (within-word, between-words) were then constructed by reduplicating metric values in accordance with their raw frequencies for each type.

The culmination of this process was 36 (4 languages X 3 transition conditions X 3 metric values) pairs of within-word and between-word distributions. The following section provides a summary of the distributions and reports the separability of transition types for each distribution pair.

5 Results

5.1 Summary of distributions

Table 3 provides a summary of the distributions. Due to TP values ranging zero to one, all TP values are log-scaled. This scaling also ensured a more normal distribution of TP values.

Table 3: Summary of metric distributions for within-word (*ww*) and between-word (*bw*) transitions by mean (μ) and standard deviation (σ) for each language corpus

		English		Cantonese		Japanese		Arabic	
		bw	*ww*	*bw*	*ww*	*bw*	*ww*	*bw*	*ww*
Diphone[1↔1]	FTP $\mu =$	-3.531	-2.765	-2.871	-2.037	-3.141	-1.906	-2.94	-2.278
	$\sigma =$	1.027	1.006	0.815	0.996	1.106	1.222	0.990	0.862
	BTP $\mu =$	-3.529	-2.766	-2.994	-1.970	-2.448	-2.207	-2.802	-2.325
	$\sigma =$	1.079	0.995	0.921	1.036	1.135	0.990	0.906	0.837
	MI $\mu =$	0.044	0.753	0.334	1.255	0.419	1.126	-0.011	0.459
	$\sigma =$	0.802	0.798	0.677	0.834	0.887	0.964	0.791	0.720
Triphone[2↔1]	FTP $\mu =$	-3.348	-2.236	-2.739	-1.577	-2.893	-1.897	-2.749	-1.502
	$\sigma =$	1.083	1.211	0.860	1.100	1.281	1.209	1.157	1.048
	BTP $\mu =$	-6.092	-5.017	-4.731	-3.778	-4.386	-4.109	-4.856	-3.812
	$\sigma =$	1.459	1.575	1.353	1.466	1.562	1.426	1.290	1.301
	MI $\mu =$	0.227	1.341	0.739	2.566	0.667	1.325	0.181	1.267
	$\sigma =$	0.953	1.219	0.746	1.014	1.117	1.106	1.083	1.117
Triphone[1↔ 2]	FTP $\mu =$	-6.027	-5.015	-4.775	-3.761	-4.464	-4.093	-4.823	-3.751
	$\sigma =$	1.393	1.576	1.311	1.471	1.663	1.428	1.466	1.314
	BTP $\mu =$	-3.326	-2.219	-2.850	-1.733	-2.262	-1.924	-2.536	-1.511
	$\sigma =$	1.095	1.231	0.919	1.076	1.159	1.051	1.143	1.074
	MI $\mu =$	0.247	1.360	0.478	1.562	0.605	1.313	0.255	1.319
	$\sigma =$	0.925	1.234	0.746	0.988	1.010	1.086	1.092	1.137

Comparing metric means, we see that within-word transitions are more probable, or share more information, than between-word transitions without exception. This provides empirical support of the generalization originally stated by Saffran et al. (1996a). Further, the cross-linguistic occurrence of this pattern, in spite of the variablility of languages, supports the notion that separability of transition types could function as a means of segmenting speech. However, distributions do overlap (seen through incorporating standard deviation), meaning this information alone does not allow perfect segmentation. This imperfect segmenting provides empirical support of authors such as

Johnson and Tyler (2010) who have questioned the effectiveness of statistical learning based speech segmentation with natural language.

Each pair of distributions is also visualizable via histograms and corresponding density plots. Figure 2 demonstrates the separability of between-word and within-word transitions for each metric in the triphone[2↔1] transition condition for the HKCC. This graph was chosen due to particularly clear separability of transition types. TP values closer to zero correspond to higher probabilities due to the log-scaling. Visual inspection of these graphs also corroborates the notion that within-word transitions have higher TP values and share more information.

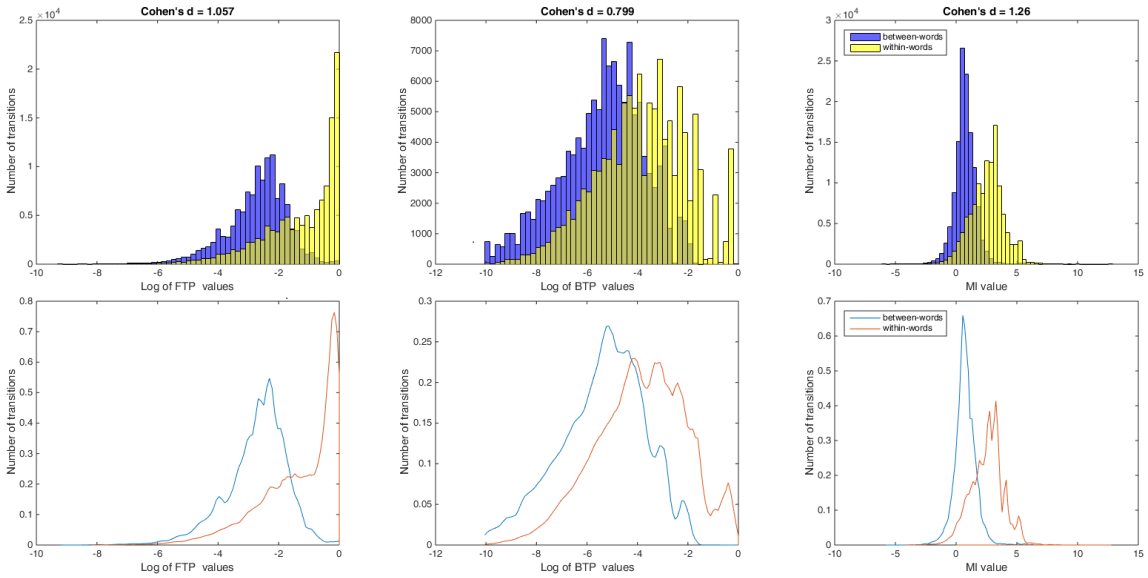

Figure 2: Histogram and density plot comparisons of metric distributions for triphone[2↔1] transitions in the HKCC. The less overlap of distributions, the more separability of within-word and between-word transition types.

5.2 Significance tests and effect sizes

To assess whether metric distributions allow for significant separability of transition type, a two-sample t-test on the distributions was performed. Prior to this, a two-sample F-test for equal variance was performed to determine if Student's t-test for equal variance or Welch's t-test for unequal variance was appropriate. Finally, to compare the effect size of separability that each metric affords, Cohen's *d* was calculated. The results are reported in Table 4.

Table 4: Summary of within-word and between-word separability by metric distributions. All t-tests were significant ($p < 0.001$); thus p-values are not reported. The highest average for each language is highlighted in blue.

		Transition Condition						
		Diphone[1↔1]		Triphone[2↔1]		Triphone[1↔2]		
		t-value	Cohen's d	t-value	Cohen's d	t-value	Cohen's d	$\mu(d)$
English:	FTP	325.91	0.712	373.23	0.869	262.70	0.643	0.741
	BTP	315.25	0.703	273.02	0.667	366.53	0.856	0.742
	MI	384.11	0.818	393.71	0.904	395.06	0.905	0.876
Cantonese:	FTP	268.62	0.819	290.19	1.017	179.70	0.685	0.840
	BTP	302.96	0.922	166.82	0.640	275.54	0.976	0.846
	MI	356.47	1.027	341.87	1.143	305.05	1.055	1.075
Japanese:	FTP	600.57	0.938	402.63	0.747	119.78	0.241	0.642
	BTP	122.47	0.231	93.17	0.186	153.21	0.304	0.240
	MI	431.61	0.710	299.23	0.569	343.60	0.638	0.639
Arabic:	FTP	159.82	0.704	244.14	1.012	166.03	0.737	0.818
	BTP	124.10	0.542	177.42	0.753	200.96	0.856	0.717
	MI	140.59	0.613	218.16	0.892	211.13	0.866	0.790
	$\mu(d)$		0.728		0.783		0.730	

All F-tests had $p < .006$ and t-tests had $p < 0.001$; thus individual p-values are not reported. Every distribution comparison failed the F-test for equal variance; therefore, the subsequent test used to check for reliably different means was Welch's T-Test which is more conservative and assumes unequal variance. The uniformity of unequal variance here is not particularly surprising given there are unequal numbers of within-word and between-word transitions in language. The reliable separability of metric means was confirmed for all cases by the t-tests. To gauge the overall separability afforded by each metric, Cohen's d was calculated as a measure of effect size. A Cohen's $d > 0.8$ is regularly considered a large effect size with increasing values interpretable as more separability afforded. This correspondence between a larger Cohen's d and more separability is clearly seen in the visualization of distributions for diphone transitions in the CSJ, shown in Figure 3.

Concretely, Figure 3 shows that, of the diphone transition condition in Japanese, FTP provides the most separability of within-word and between-word transition types and BTP provides the least. Similarly, looking back to Figure 2, MI affords the most separability of transition type which is confirmed by the largest Cohen's d in the Cantonese triphone[2↔1] transition condition. The density plots for all languages are provided on the final page of this paper for the readers visual inspection.

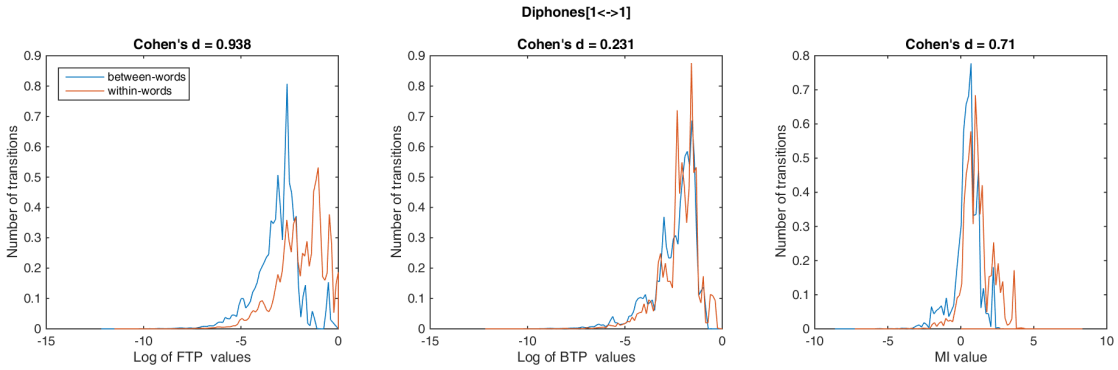

Figure 3: Density plot comparisons and Cohen's d values of metric distributions for diphone[1↔1] transitions in the CSJ. The less overlap of distributions, the more separability of transition types and a higher Cohen's d value.

5.3 Results discussion

Comparing Cohen's d values across metrics and transition conditions facilitates the identification of both cross-linguistic and language-specific patterns. In Table 4, we see that MI, on average, provides the most separability of transition types across all conditions and all languages (MI $\mu =$ 0.845; FTP $\mu = 0.760$; BTP $\mu = 0.636$). Mutual Information, however, is not consistently the most informative language-by-language. For English and Cantonese, MI affords the most separability; for Japanese and Arabic, FTP affords the most separability. This lines up with the prediction that the former two languages would pattern together and the latter two would pattern together given their similar phonemic inventory composition. The fact that MI is the most informative in English also aligns well with previous corpus research which showed empirically that MI affords better speech segmentation than other metrics (Swingley 1999).

Considering transition condition now, we see that the triphone[2↔1] condition provides the most separability of transition types as measured by averaging Cohen's d across all metrics (diphone[1↔1] $\mu = .728$; triphone[2↔1] $\mu = 0.783$; triphone[1↔2] $\mu = 0.730$). This patterns is true for all languages except Japanese, in which diphones provide the most separability. This is likely a result of the restricted phonotactics in Japanese as compared to the other languages.

There is also a consistent pattern of FTP outperforming BTP in transition condition triphone[2↔1] and falling behind BTP in transition condition triphone[1↔2]. This is expected as FTP incorporates diphone frequency in triphone[2↔1] but BTP incorporates only monophone frequency. This pattern is reversed in the triphone[1↔2] transition condition where BTP outperforms FTP. Thus, this is the result of FTP and BTP each having an advantageous transition condition (triphone[2↔1] and triphone[1↔2], respectively). A comparison of FTP and BTP across languages shows that their relative informativeness is language-specific. In English, they appear to have similar informativeness as they have similar values in the diphone[1↔1] condition (FTP $d = 0.712$, BTP $d = 0.704$) and similar values in their respective advantageous condition (FTP $d = 0.869$, BTP $d =$ 0.857). In Cantonese, BTP appears to be more informative than FTP as seen in the diphone[1↔1] condition (FTP $d = 0.450$, BTP $d = 1.222$) and contrasting values in their respective advantageous condition (FTP $d = 1.057$, BTP $d = 1.270$). In Japanese and Tunisian Arabic, FTP appears to be more informative than BTP via the same logic.

Comparing languages, we see that Japanese, on average, has the least separability of transition

types while Cantonese has the most. This is consistent with the predictions that Japanese, as a morphologically rich language, will have less separability than other languages, and that Cantonese, as a tone language, will have more. Thus, all three predictions made based on language variability were confirmed.

6 General discussion

In combination, these results provide strong support that there is statistical information in natural speech cross-linguistically which *could* be leveraged for speech segmentation by infants. For all transition conditions and all metrics, there was significant separability between transition types. However, it is important to note that the within-word and between-word transitions are not perfectly separable in any case. This observation is in line with previous research (e.g. Cairns et al. 1997) and supports the argument that other cues, such as stress patterns, must also be used by infants for more accurate speech segmentation (e.g. Yang 2004). Nonetheless, the patterns herein suggest that statistical learning could enable the delineation of words in continuous speech to some extent. That is to say, provided infants are able to track transitional patterns (as has been assumed by others, e.g. Daland and Pierrehumbert 2011), they are likely able to accurately identify some words. This is consistent with the idea that infants first learn to segment a small set words via statistical learning and then develop other segmentation strategies, such as the use of stress, based on commonalities in the set of now known words (Mattys et al. 2005; Thiessen and Saffran 2003).

It is important to state that the significance of this work relies on infants developing, or having innately, knowledge of speech sounds and possessing a robust statistical learning mechanism capable of tracking transitions between such sounds on a large-scale. Further, these results were achieved with perfect word-boundary knowledge. This raises the question of how an infant would begin to develop distributions analogous to the ones here (see Cairns et al. 1997:for more discussion) in the first place. Daland and Pierrehumbert (2011) address this question directly by suggesting that between-word transitions marked with acoustic cues (e.g. pauses between phrases) can be used to bootstrap other between-word transitions. A combination of their proposal and the current project may also be fruitful in the future.

As a final thought, the kind of separability found here may also find use in language typology. These analyses provide another dimension on which to categorize languages and such categorization may provide insight into other problems in the future.

7 Conclusion

This work investigated the statistical separability of within-word and between-word segment transitions in four languages. Three metric types (FTP, BTP, MI) that have been reported in the literature as ways of encapsulating statistical regularities in language were applied to segment transitions resulting in distributions of each metric for each transition type. Three transition conditions (diphone[1↔1], triphone[2↔1], triphone[1↔2]) were tested in line with previous speech segmentation and speech processing research. Both t-tests and Cohen's *d* calculations showed separability of transition types for all metrics, for all transition conditions and for all languages. Mutual Information was consistently found to be the most informative metric. Further, the triphone[2↔1] transition condition provided the most separability of transition types. Japanese afforded the least separability of transition types and Cantonese afforded the most. Finally, English and Cantonese

patterned similarly and Tunisian Arabic and Japanese patterned similarly in terms of separability afforded by each metric. This patterning aligned well with each languages phonemic inventory composition.

8 Acknowledgements

I would like to acknowledge my classmates Michael McAuliffe, Blake Allen and Zoe Lam for their assistance in helping me complete this project and thank professors Dr. Babel, Dr. Currie-Hall, Dr. Stemberger and Dr. Vatikiotis-Bateson for their guidance.

References

Bauer, R. S. and Benedict, P. K. (1997). Modern cantonese phonology. In Winter, W., editor, *Trends in Linguistics: Studies and Monographs 102*. Mouton de Gruyter, Berlin · New York.

Boersma, P. and Weenink, D. (2014). Praat: doing phonetics by computer [computer program].

Brent, M. (1999a). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105.

Brent, M. (1999b). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Science*, 3(8):294–301.

Cairns, P., Shillcock, R. C., Chater, N., and Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, 33:111–153.

Cole, R. A. and Jakimik, J. (1980). A model of speech perception. In Cole, R. A., editor, *Perception and production of fluent speech*, pages 133–163. Erlbaum, Hillsdale, USA.

Daland, R. (2009). *Word Segmentation, Word Recognition, and Word Learning: A Computational Model of First Language Acquisition*. PhD thesis, Northwestern University.

Daland, R. and Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Science*, 35:119–155.

de la Cruz Pavía, I. (2012). *Chunking the input: on the role of frequency and prosody in the segmentation strategies of adult bilinguals*. PhD thesis, Universidad del País Vasco.

Gibson, M. (1998). *Dialect Contact in Tunisian Arabic: sociolinguistic and structural aspects*. PhD thesis.

Glass, J. R. (2003). A probabilistic framework for segment-based speech recognition. *Computer Speech & Language*, 17:137–152.

Hall, K. C., Allen, B., Fry, M., Mackie, S., and McAuliffe, M. (2015). Phonological corpustools [computer program].

Ito, J. and Mester, R. A. (1995). Japanese phonology. In Goldsmith, J. A., editor, *The Handbook of Phonological Theory. Blackwell Handsbook in Linguistics*. Blackwell Publishers.

Johnson, E. K. and Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44:548–567.

Johnson, E. K. and Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, 13(2):339–345.

Lehiste, I. (1960). An acoustic-phonetic study of open juncture. *Phonetica, Supplementurum ad*, 5:1–54.

Leung, M. T. and Law, S. P. (2001). Hkcac: The hong kong cantonese adult language corpus. *International Journal of Corpus Linguistics*, 6:303–326.

Maekawa, K. (2003). Corpus of spontaneous japanese: Its design and evaluation. Tokyo.

Masmoudi, A., Khmekhem, M. E., Esteve, Y., Belguith, L. H., and Habash, N. (2014). A corpus and phonetic dictionary for tunisian arabic speech recognition. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mattys, S. K., White, L., and Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, 130(4):477–500.

Pellegrino, F., Coupé, C., and Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, 87:3:539–558.

Perruchet, P. and Desaulty, S. (2008). A role for backward transitoinal probabilities in word segmentation. *Memory and Cognition*, 36(7):1299–1305.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). Buckeye corpus of conversation speech (2nd release). Ohio State Univeristy.

Roach, P. (2009). *English Phonetic and Phonology: A Practical Course, 4th Ed.* Cambrige University Press, Cambridge.

Rytting, C. A. (2004). Segment predictability as a cue in word segmentation: Application to modern greek.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996b). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, 35:606–621.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27.

Swingley, D. (1999). Conditional probability and word discovery: A corpus analysis of speech to infants. In Hahn, M. and Stoness, S. C., editors, *Proceedings of the 21st annual conference of the cognitive science society*, pages 724–729, Mahwah, NJ: LEA.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50:86–132.

Thiessen, E. and Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month old infants. *Developmental Psychology*, 39:706–716.

Toro, J. M., Sinnett, S., and Soto-Faraco, S. (2005). Speech segmentation by statistical learning

depends on attention. *Cognition*, 97:B25–B34.

Yang, C. D. (2004). Universal grammar, statistics, or both? *TRENDS in Cognitive Science*, 8:451–456.

Density Plots

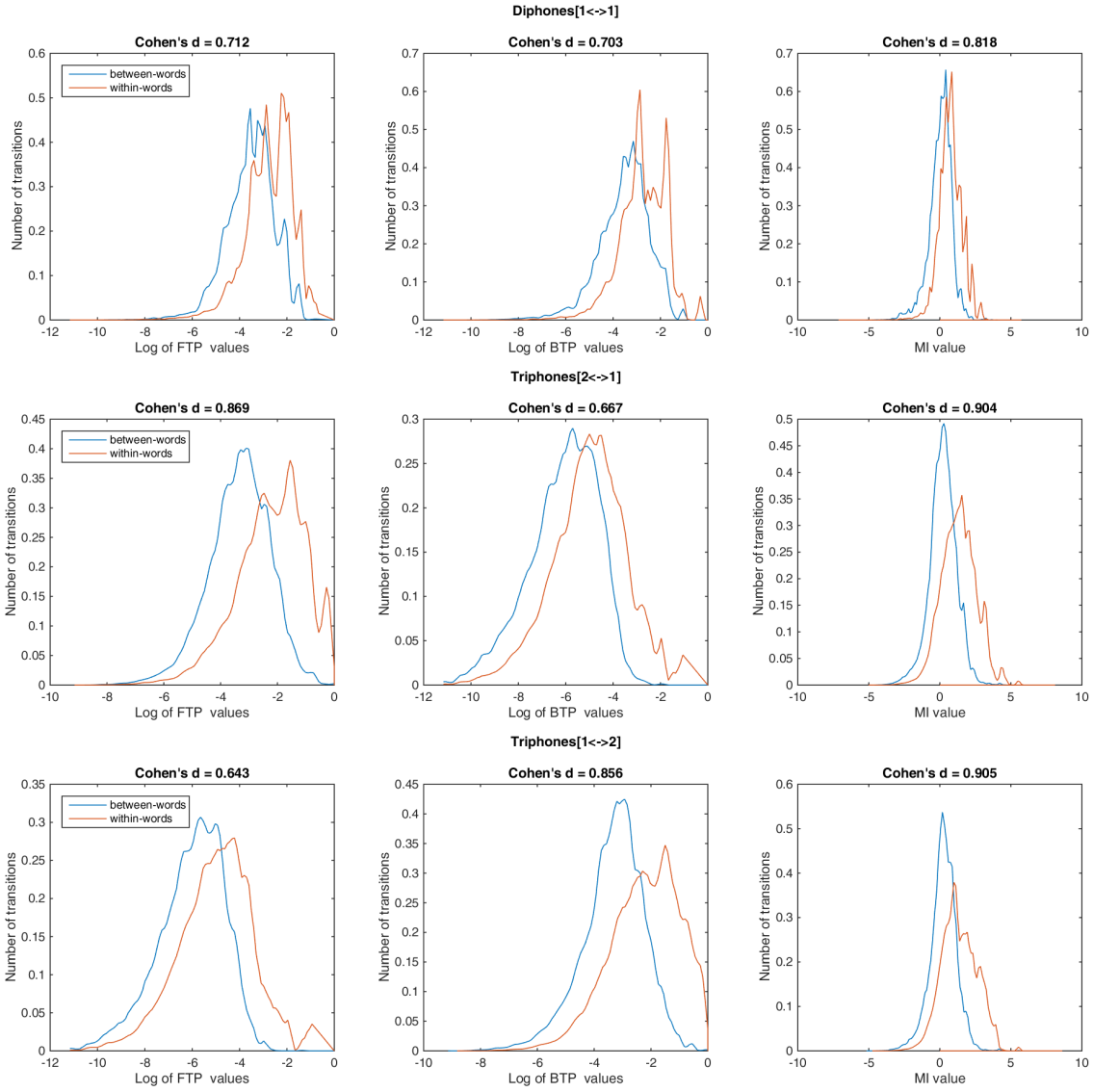

Figure 4: English density plots

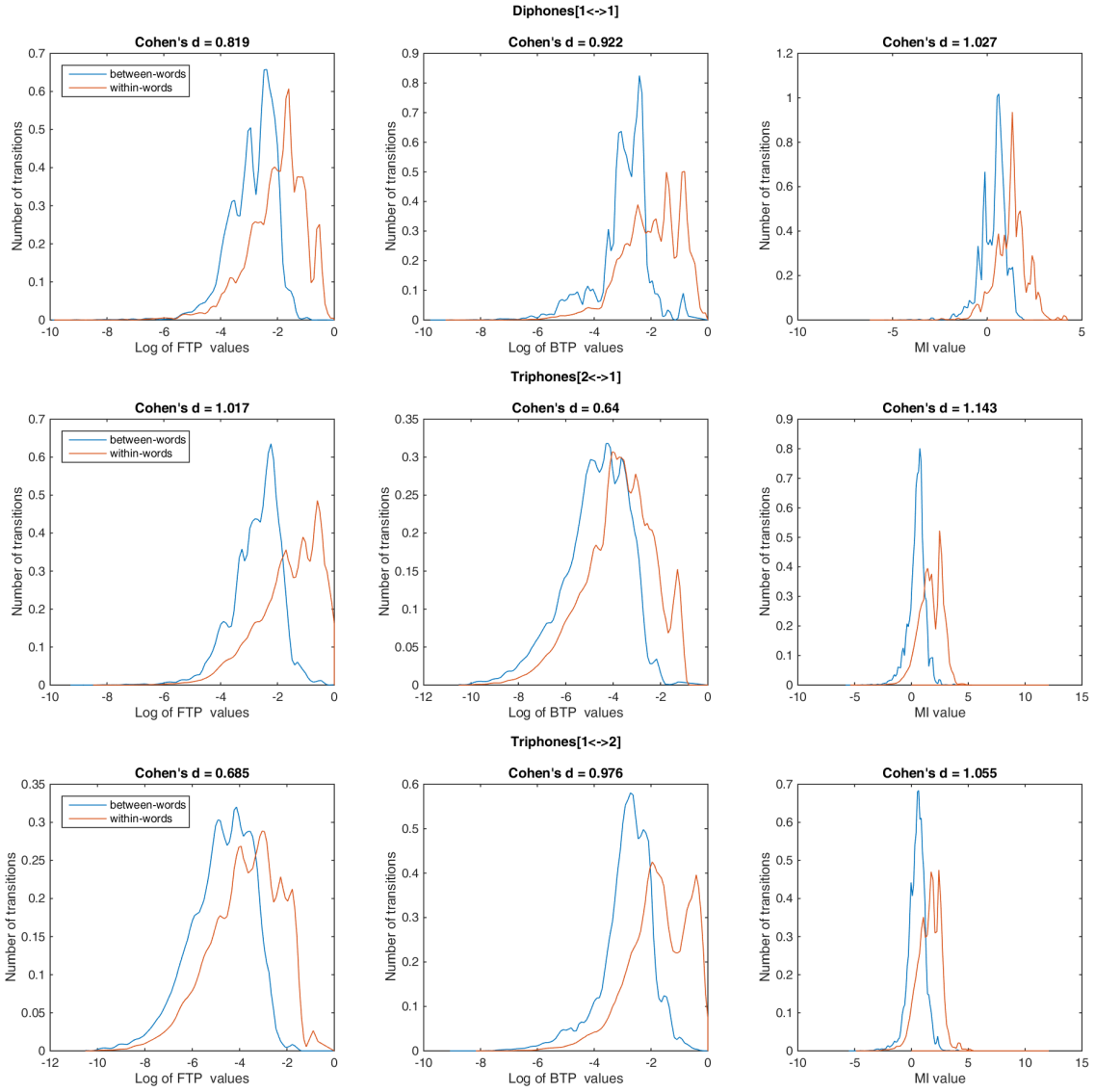

Figure 5: Cantonese density plots

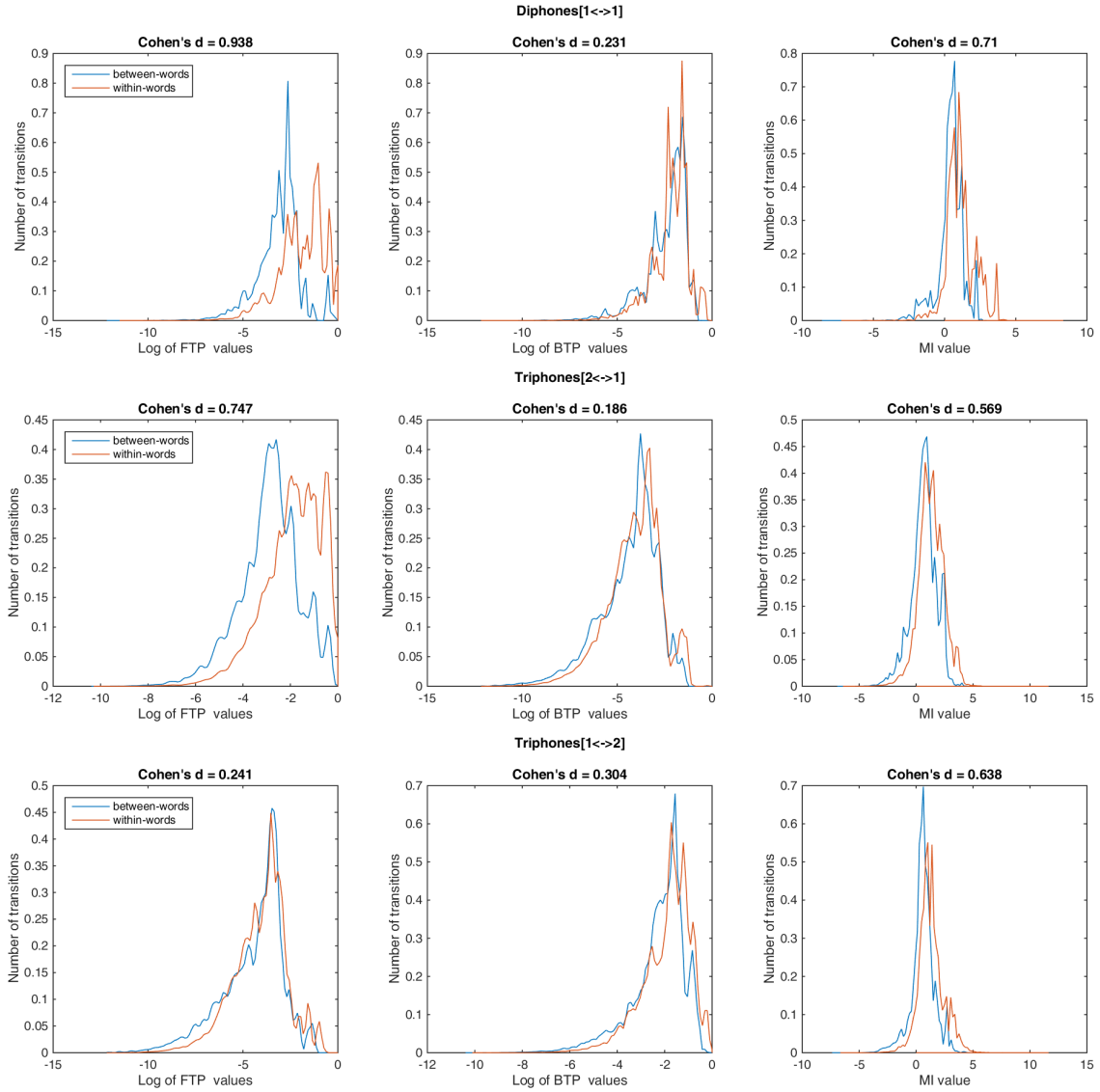

Figure 6: Japanese density plots

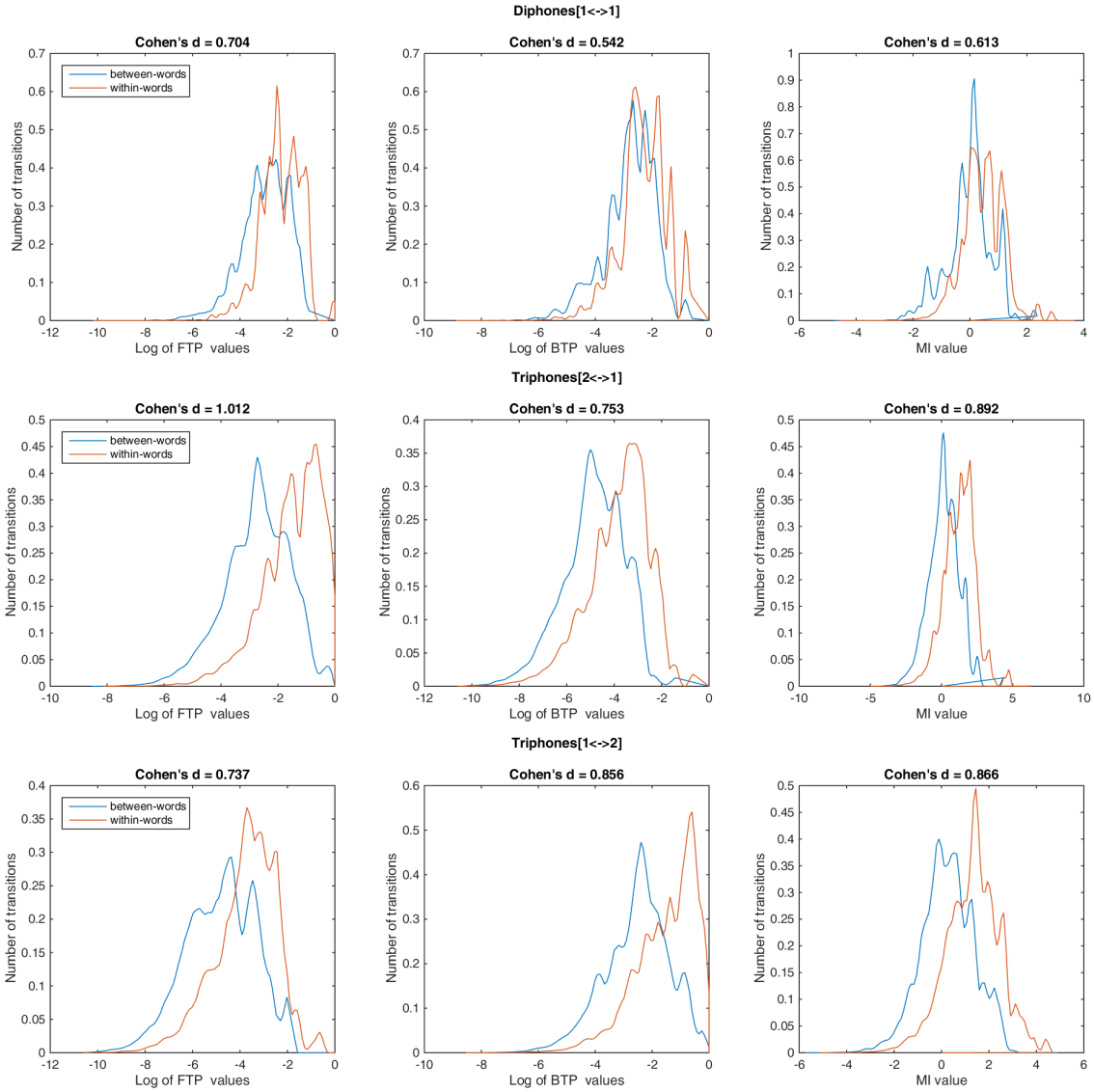

Figure 7: Tunisian Arabic density plots