STATISTICAL DIGRAM AND TRIGRAM ANALYSIS OF TURKISH IN TERMS OF COVERAGE AND ENTROPY FOR POSSIBLE LANGUAGE AND SPEECH BASED APPLICATIONS

İbrahim Baran Uslu [1], *Asım Egemen Yılmaz* [2], *Hakkı Gökhan İlk* [2]

[1] Başkent University, Faculty of Engineering, Electrical and Electronics Engineering Department
Eskişehir Yolu 20.km Etimesgut 06530 Ankara-TURKEY
phone: +90-312-2341010, fax: +90-312-2341051, email: ibuslu@baskent.edu.tr

[2] Ankara University, Faculty of Engineering, Electronics Engineering Department
Tandoğan 06100 Ankara-TURKEY
phone: +90-312-2033305, fax: +90-312-2125480, aeyilmaz@eng.ankara.edu.tr, ilk@ieee.org

ABSTRACT

In this study two frameworks, made up of digrams and trigrams, are built for a complete coverage of the Turkish language. In addition, character, digram and trigram entropy values for Turkish, English and Spanish are compared. Examining meaningful Turkish texts, we have achieved the result that, there are 3 major digram clusters which constitute slightly more than 60% of Turkish texts. Similar to digram distributions, there are 3 major trigram clusters which cover almost 40% of Turkish texts. The statistics show that, for 99% coverage of Turkish, 391 (of 841 theoretical) digrams and 3,396 (of 24,389 theoretical) trigrams are sufficient. The results of this study would constitute a general roadmap for rapid coverage to researchers who would like to work on Turkish language and speech based applications. As an application, the results could lead to a general framework for setting up the rules of prioritization in duration modeling in concatenative text-to-speech synthesis systems.

1. INTRODUCTION

Converting written text into oral form, that is automatic reading of texts, has been a very interesting and exciting area for researchers from various disciplines. In the last decades, the research has been focused on concatenative text-to-speech (TTS) synthesis systems, which are based on combining pre-recorded speech segments via some specific signal processing techniques. Interested readers could proceed to [1-3] for detailed information related to these methods. The question of interest regarding concatenative speech synthesis is the units (i.e. speech segments) to be used in concatenation. Various segments of different lengths, ranging from phonemes up to words and phrases, can be used for this purpose. Certainly, there is a tradeoff between the segment size and the signal processing work load. In case that the phonemes are used, the number of operations during concatenation will be much more than the case of using diphones. On the other hand, as the speech segment gets bigger (such as syllable, triphone, word, etc.), the number of total combinations will grow exponentially; which would dramatically increase the memory requirements eventually. For this reason, diphones are frequently preferred in the recent studies, since there is a convenient number of diphones (about 1600 in Turkish). Moreover, they include the natural transition from the middle of the first phoneme to the middle of the second phoneme.

In modern standard Turkish, there are 29 letters and 44 phonemes, which are listed in Table I [4]. Some of the letters have variations depending on the articulation and the context of usage. Besides this, some phonemes can also be lengthened, palatalized, or both; according to where they are used in words depending on the adjacent sounds.

In concatenative synthesis, after deciding on the speech segment type, signal processing methods for concatenation take place. Here, the basic method is “overlap and add”. In this technique, periods from the end of the first and the beginning of the second unit are selected. After windowing, these units are overlapped and finally added. Independent of the overlap and add method used in concatenative speech synthesis, the unit selected has utmost importance. Distortions during synthesis increase due to misselected units. Moreover, even if the selected units are “proper”, there is a significant challenge while picking-up the sub-units. Hence, for a natural speech synthesis, the system should be able to employ the “proper” units, sub-units, duration and later on intonation models. In light of this information, for a natural speech synthesis, the prerequisite is to establish the coverage statistics and the corresponding entropy result for that specific language.

 ISSN 2076-1465

TABLE I
LETTERS AND PHONEMES IN TURKISH

Letter	Phoneme (IPA)	Example	Letter	Phoneme (IPA)	Example	Letter	Phoneme (IPA)	Example
a	ɑ	**ɑ**nı (memory)	j	ʒ	müʒde (surprise)	s	s	**s**es (sound)
	a	l**a**f (utterance)	k	c	**c**edi (cat)	ş	ʃ	aʃı (vaccine)
b	b	**b**al (honey)		k	a**k**ıl (mind)	t	t	üt**t**ü (iron)
c	ʤ	ʤam (glass)	l	l	l**a**le (tulip)	u	ʊ	kʊlak (ear)
ç	ʧ	se**ʧ**im (selection)		ɬ	ku**ɬ** (villein)		u	**u**ğur (fortune)
d	d	**d**e**d**e (grandfather)	m	m	da**m** (roof)	ü	ʏ	ʏmit (expectance)
e	ɛ	der**ɛ** (river)	n	n	a**n**ı (memory)		y	d**y**ğme (button)
	e	**e**lma (apple)		ŋ	sü**ŋ**gü (bayonet)	v	v	**v**ar (present)
f	f	**f**asıl (trial)	o	ɔ	sɔru (question)		ʊ	taʊuk (chicken)
g	ɟ	ɟenç (young)		o	**o**ğlak (goat)	y	j	**j**at (yacht)
	g	kar**g**a (crow)	ö	œ	œrtü (tablecloth)		ːɪ	huːɪ (habit)
ğ		ya**ğ**mur (rain)		ø	øğren (learn)	z	z	a**z**ık (viaticum)
h	h	**h**asta (patient)	p	p	i**p** (rope)		ʐ	yoʐ (virgin)
ı	ï	**ï**sï (heat)	r	r	**r**af (shelf)			
i	i	**i**ğde (oleaster)		ɾ	**ɪ**fmak (river)			
	ɪ	sɪmɪt (bagel)		ɣ	biɣ (one)			

The ideal approach for computation of such statistics would be based on data extraction from pre-recorded daily life dialogues of native speakers; where the distribution of these speakers should better be homogenous in terms of the factors such as gender, age, education level, dialect, etc., as implemented by Salor *et al* [5]. Certainly, there would be some practical issues and hurdles at various steps of this process, such as suppression and elimination of background noise, proper unit identification and labeling, etc. For that reason, our approach in this study is; computation of exact written language statistics from literal works, which will be succeeded by retrieval of approximate statistical data for spoken language. Discussions regarding the granularity of such an approximation will be held in the upcoming sections.

To our belief, the results of this study might provide bases for various purposes. The approximate spoken language statistics provided in this study might construct a guideline to researchers, who will deal with diphone/ triphone based concatenative speech synthesis. These results would give an idea for prioritization of diphones/ triphones for rapid language coverage with minimum unit-recording effort. Another possible outcome of these results might be ambiguity resolution in speech recognition applications. Regardless of these potential applications, our main aim is to get advantage of these results while constructing a general framework of a rule-based duration model.

The outline of this paper is as follows. After this introductory section, in Section 2, we explain the digram and trigram frameworks in detail. After the classification of the digrams and trigrams, we give the statistical results obtained from the examined Turkish e-books. In Section 3, we discuss on entropy of Turkish while comparing with the languages English and Spanish. And in the last section we discuss the potential uses of the results of this study for possible speech and language based applications.

2. COVERAGE STATISTICS OF TURKISH IN TERMS OF DIGRAMS AND TRIGRAMS

2.1 Motivation

29 letters; the entire Turkish alphabet (**A**) is compromised of vowels (**V**) and consonants (**C**), where;

V = {a, e, ı, i, o, ö, u, ü}

C = {b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z}

The consonants can further be grouped into 4 subgroups, according to their hardness/ softness and sustainability/ unsustainability properties, as follows:

$\mathbf{C_{soft} = C_s = \{b, c, d, g, ğ, j, l, m, n, r, v, y, z\}}$
(the set of soft consonants)

$\mathbf{C_{hard} = C_h = \{ç, f, h, k, p, s, ş, t\}}$ *(the set of hard consonants)*

$\mathbf{C^{sustainable} = C^s = \{f, ğ, h, j, l, m, n, r, s, ş, v, y, z\}}$
(the set of sustainable consonants)

$\mathbf{C^{unsustainable} = C^u = \{b, c, ç, d, g, k, p, t\}}$
(the set of unsustainable consonants)

$\mathbf{C_h^s = C_h \cap C^s = \{f, h, s, ş\}}$
(the set of hard and sustainable consonants)

$\mathbf{C_h^u = C_h \cap C^u = \{ç, k, p, t\}}$
(the set of hard and unsustainable consonants)

$\mathbf{C_s^s = C_s \cap C^s = \{ğ, j, l, m, n, r, v, y, z\}}$
(the set of soft and sustainable consonants)

$\mathbf{C_s^u = C_s \cap C^u = \{b, c, d, g\}}$
(the set of soft and unsustainable consonants)

For the following sections, $\mathbf{v \in V}$, $\mathbf{c_x^y \in C_x^y}$; where bold **v** and $\mathbf{c_x^y}$ are elements of the sets **V** and $\mathbf{C_x^y}$, respectively. It should be noted that, with such a notation, we can identify the plosive consonants: {b, d, g, p, t, k}, which is a subset of the union of the sets $\mathbf{C_h^u}$ and $\mathbf{C_s^u}$. In other words, at the instances where the elements of this set are encountered, it is possible to identify the potential candidates of the loss of "plosion" effect. Similarly, all vowels are prone to unexpected shortening/lengthening effect in synthesis. Other sets and their unions/intersections might similarly be correlated to some other misleading effects of synthesis. Hence, the starting point of our study, is nothing but the statistical data analysis about the occurrence rate of all letter groups (i.e. n-grams, especially for n = 1, 2 and 3) for meaningful Turkish texts. The results of this analysis will give an overall idea about the expected occurrence rates of the relevant misleading effects of synthesis.

2.2 Digram Statistics

Digrams or bigrams are taken as groups of two letters, in this study; different from the n-gram word groups reported in [6] and syllable bigrams reported in [7]. There exist 29x29 = 841 possible digrams in Turkish. Regarding our notation and classification, since the Turkish alphabet can be grouped into 5 major subsets (i.e. **V**, $\mathbf{C_h^s}$, $\mathbf{C_h^u}$, $\mathbf{C_s^s}$, $\mathbf{C_s^u}$), these digrams can be grouped into 5x5 = 25 main clusters (classes of digrams such as: $\mathbf{V\text{-}C_s^s}$, $\mathbf{C_s^s\text{-}V}$, $\mathbf{C_s^u\text{-}V}$) some examples of which can be seen in Table II, where x is any letter. Using this classification, we examined some Turkish e-books (with non-technical but literal content, properties of which are given in Table III) in terms of digram distributions. The statistics we obtained can be seen in Table IV. The results show that $\mathbf{(V\text{-}C_s^s)}$ and $\mathbf{(C_s^s\text{-}V)}$ clusters constitute almost 50% of Turkish digrams. Adding the $\mathbf{(C_s^u\text{-}V)}$ cluster to these two, coverage of 60% of the written language is achieved. Furthermore, for 99% coverage, the required number of the digrams is 391 of theoretical 841. This shows us that, some digram combinations are very much rarely used in Turkish texts. Since Turkish can be considered as a "phonetic language", except some specific words adopted from foreign languages, such as Arabic, Persian and French (i.e. grapheme-to-phoneme mapping is one to one in most cases), from the distributions of the digrams, we obtained the corresponding number of diphones. The calculation is done by multiplying each letter with its possible number of phoneme variants, making the worst case assumption: *each letter might correspond to all of its relevant phonemes.* There will be 1,076 diphones of 1,936 theoretical, needed for 99% coverage of spoken Turkish. The digram and diphone coverage curves are plotted in Figure 1.

TABLE II
SOME EXAMPLES FOR THE DEFINED DIGRAM CLUSTERS

1. $\mathbf{x(c_s^s)(c_h^s)x}$: e**ms**al (example)	**9.** $\mathbf{x(c_s^u)(c_s^u)x}$: e**cd**ad (ancestor)	**18.** $\mathbf{x(c_h^u)(c_s^s)x}$: ka**çm**a (escape)
2. $\mathbf{x(c_s^s)(c_h^u)x}$: i**mk**an (facility)	**10.** $\mathbf{x(c_s^u)(v)x}$: san**cı** (pain)	**19.** $\mathbf{x(c_h^u)(c_s^u)x}$: i**kb**al (wish)
3. $\mathbf{x(c_s^s)\text{-}(c_s^s)x}$: a**nl**a (do understand)	**11.** $\mathbf{x(c_h^s)(c_h^s)x}$: müe**ss**ese (establishment)	**20.** $\mathbf{x(c_h^u)(v)x}$: kitap**çı** (book seller)
4. $\mathbf{x(c_s^s)(c_s^u)x}$: a**mc**a (uncle)	**12.** $\mathbf{x(c_h^s)(c_h^u)x}$: a**şç**ı (cook)	**21.** $\mathbf{x(v)(c_h^s)x}$: **ar**aba (car)
5. $\mathbf{x(c_s^s)(v)x}$: **ma**karna (pasta)	**13.** $\mathbf{x(c_h^s)(c_s^s)x}$: i**hm**al (negligence)	**22.** $\mathbf{x(v)(c_h^u)x}$: p**at**ates (potato)
6. $\mathbf{x(c_s^u)(c_h^s)x}$: a**ds**ız (without name)	**14.** $\mathbf{x(c_h^s)(c_s^u)x}$: me**şg**ul (busy)	**23.** $\mathbf{x(v)(c_s^s)x}$: k**ol** (arm)
7. $\mathbf{x(c_s^u)(c_h^u)x}$: -	**15.** $\mathbf{x(c_h^s)(v)x}$: **ha**tıra (memory)	**24.** $\mathbf{x(v)(c_s^u)x}$: d**ed**e (grandfather)
8. $\mathbf{x(c_s^u)(c_s^s)x}$: a**bl**a (big sister)	**16.** $\mathbf{x(c_h^u)(c_h^s)x}$: e**ks**en (axis)	**25.** $\mathbf{x(v)(v)x}$: s**aa**t (watch)
	17. $\mathbf{x(c_h^u)(c_h^u)x}$: teşe**kk**ür (thanks)	

TABLE III
PROPERTIES OF THE EXAMINED E-BOOKS

Title of the e-book	Author	Type	Number of pages	.txt File size (kB)	Number of words	Number of digrams
Veda Yemeği *"Farewell Dinner"*	Michel TOURNIER	novel (translation)	111	385	52,537	274,152
Son Antlaşma *"The Last Treaty"*	Can ERYÜMLÜ	novel	250	708	95,988	479,674
Bir Hasta Sahibinin Hastane Günlüğü *"The Hospital Diary of a Patient Owner"*	Doğan PAZARCIKLI	diary	40	122	16,472	84,021

TABLE IV

DIGRAM DISTRIBUTIONS

Cluster	E-book1	E-book2	E-book3
x(V–C_s^s)x	**28.15%**	**28.40 %**	**28.72 %**
x(C_s^s–V)x	**21.66%**	**21.95 %**	**21.60 %**
x(C_s^u–V)x	**11.15%**	**11.54 %**	**11.18 %**
x(C_h^u–V)x	7.38 %	6.94 %	7.29 %
x(V–C_h^u)x	5.67%	5.35 %	5.73 %
x(C_h^s–V)x	5.15 %	5.10 %	5.90 %
x(C_s^s–C_s^s)x	3.84 %	4.17 %	3.52 %
x(V–C_h^s)x	3.96%	4.16 %	4.33 %
x(C_s^s–C_s^u)x	3.38 %	3.11 %	2.82 %
x(V–C_s^u)x	2.38%	2.77 %	2.71 %
x(C_h^u–C_s^s)x	1.34 %	1.29 %	1.16 %

As can be seen, the curves are monotonously increasing with a decreasing slope, and they are very similar to each other; independent of the texts obtained from the e-books.

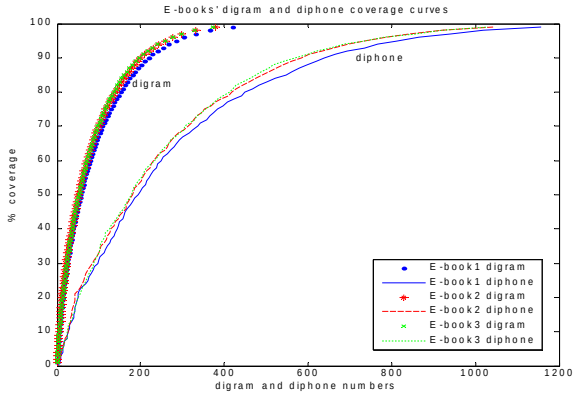

Figure 1 - Digram and corresponding diphone coverage curves of the examined e-books.

Therefore it is possible to conclude that, the digram statistics of Turkish texts is stationary. So a good point to start for recording, duration modeling, or other experiments (like intonation modeling) is the set of clusters which include **(v-c_s^s)**, **(c_s^s-v)** and **(c_s^u-v)** digram combinations for quick coverage.

2.3 Trigram Statistics

Similar to digram examinations, illustrated in the previous sub-section, we also examined the texts in terms of trigram and triphone distributions. There are 29x29x29 = 24,389 theoretical number of trigrams in Turkish. Based on our notation, we can identify 5x5x5 = 125 clusters in this case. The trigram statistics showed a distribution very much parallel to the digram statistics, such that; the most occurred trigram clusters are the extended cases of the diphone clusters, namely: **(V-C_s^s-V), (C_s^s-V-C_s^s)** and **(V-C_s^u-V)**. The percent coverage values of these three clusters are given in Table V. Since there are 125 clusters, a short list of only the most occurred trigram clusters are presented in this table. For the whole results, the readers can refer to [10]. Trigrams belonging to these three clusters cover approximately 40% of Turkish texts. Again using the worst case assumption, we calculated the required number of triphones. There will be 16,500 triphones of 85,184 theoretical, needed for 99% coverage of spoken Turkish. The coverage curves for the trigrams and the corresponding triphones can be seen in Figure 2. Trigram and triphone coverage curves are very similar to the digram and diphone coverage curves, which were given in Figure 1, again showing us the quick coverage property.

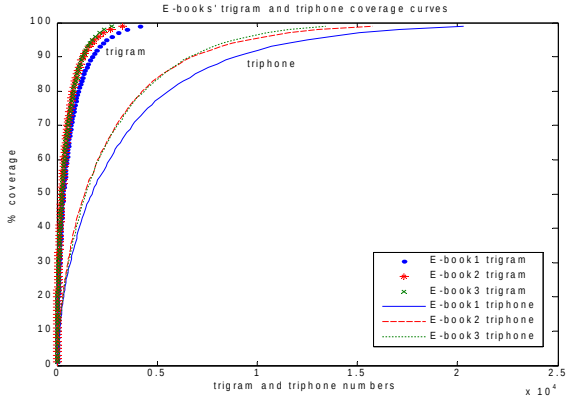

Figure 2 - Trigram and corresponding triphone coverage curves of the examined e-books.

TABLE V

TRIGRAM DISTRIBUTIONS

Cluster	E-book1	E-book2	E-book3
x(V-C_s^s-V)x	% 16.30	% 16.92	% 16.97
x(C_s^s-V-C_s^s)x	% 13.77	% 14.02	% 14.37
x(V-C_s^u-V)x	% 7.90	% 7.87	% 8.25

3. ENTROPY

Entropy; in this context, is the amount of average uncertainty content in the studied language. Therefore the written text statistics in terms of entropy may construct a guideline to researchers for applications such as ambiguity resolution in optical character recognition (OCR) and/ or approximate spoken language statistics may provide a guideline for applications in speech recognition. Unfortunately, it is not easy to calculate the entropy of a language with accuracy due to high dependency on the context of the text examined. Hence we aim to provide some basic results to report entropy values calculated over sufficient amount of text. Besides our statistical coverage analysis, we calculated the entropy of Turkish texts in terms of character entropy (F_1), digram entropy (F_2) and trigram entropy (F_3), in the unit: bits/ character, using the equations (1), (2) and (3) below [8, 9].

Character entropy:

$$F_1 = -\sum_i p(i)\log_2 p(i) \qquad (1)$$

Digram entropy:

$$F_2 = -\sum_{i,j} p(i,j)\log_2 p(j\,|\,i) = -\sum_{i,j} p(i,j)\log_2 p(i,j) + \sum_i p(i)\log_2 p(i) \qquad (2)$$

Trigram entropy:

$$F_3 = -\sum_{i,j,k} p(i,j,k)\log_2 p(k\,|\,ij) = -\sum_{i,j,k} p(i,j,k)\log_2 p(i,j,k) + \sum_{i,j} p(i,j)\log_2 p(i,j) \quad (3)$$

The entropy results for Turkish texts are given in Table VI.

TABLE VI
ENTROPY VALUES FOR THE EXAMINED E-BOOKS

Entropy (bits/ character)	E-book1	E-book2	E-book3
for character	4.38	4.36	4.37
for digram	3.88	3.75	3.76
for trigram	2.70	2.49	2.49

We can compare the calculated entropy values for Turkish with other languages such as English and Spanish. Theoretical entropy values per character (F_0) for Turkish, English [8] and Spanish [9] were calculated and presented in Table VII. In addition, statistics obtained from (1), (2) and (3) were presented in Table VII, for character (F_1), digram (F_2) and trigram (F_3) entropies for these three languages. The results are interesting in the sense that, although there is a general agreement for the average entropy values for character and digram combinations, the average entropy value for trigram for Turkish is significantly smaller than that of English and Spanish. This result is not surprising because, Turkish (possessing a large number of vowels) is an agglutinative language and consonant clusters are not permitted, except for the syllable ends, by the syllabic system rules of Turkish [4]. Therefore trigram combinations in Turkish exhibit less redundancy.

TABLE VII
ENTROPY VALUES FOR TURKISH, ENGLISH AND SPANISH

	Turkish	**English** [8]	**Spanish** [9]
F_0	4.86	4.70	5.04
F_1	4.37	4.14	4.40
F_2	3.80	3.56	3.98
F_3	2.56	3.30	3.68

4. RESULTS AND CONCLUSION

In this study, a detailed statistical analysis is performed on Turkish literal texts, by calculating the character, digram and trigram distributions, as well as, entropy of the language. The defined digram and trigram frameworks show us that, there are 3 major digram clusters covering approximately 60%, and similarly 3 major trigram clusters covering approximately 40% of the written language.

We also determined that; almost 55% of all possible diphones are sufficient for 99% coverage of spoken Turkish. Similarly, almost 20% of all possible triphones are sufficient to cover 99% of spoken Turkish. As stated before, the spoken language coverage statistics are obtained approximately from the written language statistics as Turkish is a "phonetic language".

In addition, a comparison of character, digram and trigram entropy values for Turkish, English and Spanish is performed. Even though there is a general agreement for the average entropy values for character and digram combinations, the average entropy value for trigram for Turkish is significantly smaller than that of English and Spanish. This is due to the fact that Turkish (possessing a large number of vowels) is an agglutinative language; where consonant clusters are not permitted, except for the syllable ends. Therefore trigram combinations in Turkish exhibit less redundancy. To our belief, this result may have significant impact on ambiguity resolution in OCR and other related applications. Detailed outputs of our analysis can be downloaded from the web site [10].

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Dr. Sıtkı Çağdaş İnam for his contributions.

REFERENCES

[1] T. Dutoit, *An Introduction to Text-to-Speech Synthesis,* Kluwer Academic Publishers, 1997.

[2] F. Charpentier and E. Moulines, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Proceedings of Eurospeech 89,* Paris, vol.2, pp. 13-19, 1989.

[3] B. Bozkurt and T. Dutoit, "An implementation and Evaluation of two-diphone based synthesizers for Turkish", *4th ISCA Tutorial and Research Workshop on Speech Synthesis,* pp. 247-250, Scotland, 2001.

[4] İ. Ergenç, *Spoken Language and Dictionary of Turkish Articulation,*Multilingual Publishers, 2002.

[5] Ö. Salor, et.al. , "Turkish speech corpora and recog. tools developed by porting SONIC: Towards multilingual speech recog.", *Comp. Speech and Lang,* vol. 21, pp. 580-593, 2007.

[6] Y. Çebi and G. Dalkılıç "Turkish word n-gram analyzing algorithms for a large scale Turkish corpus - TurCo", *Proc. IEEE Intern. Conf. on Information Technology: Coding and Computing (ITCC'04),* 2004.

[7] K. Günel ve R. Aşlıyan, "Turkish word error detection using syllable bigram statistics", *IEEE SIU 2006, Signal Processing and Communications Applications,* pp. 1-4, 17-19 April 2006,

[8] C.E. Shannon. "Prediction and entropy of printed English", *Bell Syst. Tech. J.,* vol. 30, pp. 47–51, January 1951.

[9] F. G. Guerrero, "On the entropy of written Spanish", submitted to IEEE Trans. On Inform. Theory, Available: http://arxiv.org/abs/0901.4784.pdf

[10] www.baskent.edu.tr/~ibuslu/FrameworksForTTS.zip