Statistical Syllable Analysis for Pronunciation Ambiguity Detection and Resolution in Text-to-Speech Synthesis Applications: A Case Study in Turkish

Ahmet Akbulut, Tugrul Adiguzel, Asim Egemen Yilmaz

Department of Electronics Engineering, Ankara University
Ankara 06100, Turkey
{aakbulut, adiguzel, aeyilmaz}@eng.ankara.edu.tr

Abstract: In this study, pronunciation ambiguity in Turkish is considered. A syllable-based ambiguity detection/resolution framework is proposed for Turkish text-to-speech synthesis applications. For this purpose, first the pronunciation ambiguity cases are identified. Such cases are classified into 7 main groups. Statistical analysis on the occurrence rate of these main groups is performed by means of the examination of meaningful Turkish texts. This first level analysis shows that especially the syllables ending with vowels (particularly with a, e and i), which are potential ambiguity sources, have significant occurrence rates. Next, the granularity of the frequency analysis is escalated to distinct syllable level. For the so-far-identified 154 exceptional syllables, the occurrence rates are computed. The results of this study will constitute a major baseline for pronunciation ambiguity detection in Turkish. The resolution of these ambiguous cases will certainly require a large lexicon. The results will also serve as a guideline for the prioritization of data inclusion to such a lexicon (i.e. lexicon enrichment) for rapid coverage. Our distinct syllable level analysis results show that by inclusion of all the words having the 100 most frequent exceptional syllables, it is possible to resolve 99% of pronunciation ambiguities in Turkish. To our belief, the findings of this study might also be applicable and useful for other languages.

Keywords: Text-to-speech synthesis; natural language processing; grapheme-to-phoneme conversion; less studied languages; pronunciation ambiguity

1 Introduction

Text-to-speech synthesis has been a popular research area with various purposes, such as increasing the ‘humanity’ in the user interactions of multimedia appliances, or aiding people with visual impairments, etc. Research studies devoted to the Indo-European linguistic family, particularly English, constitute the major portion of text-to-speech synthesis applications. Text-to-speech synthesis studies on Turkish, which started in 1990s and currently continue in academic and commercial areas, are relatively low in quantity compared to most other languages. In almost three decades, various researchers have directly or indirectly contributed to the literature regarding Turkish text-to-speech synthesis via a M. Sc. Theses [1]-[18] and Ph. D. Dissertations [19]-[20] in addition to the relevant conference proceedings [21]-[31] and journal papers [32]-[36].

For most of these publications, the general focus is on items at the signal processing level, such as the proper unit selection, concatenation, etc. Among them, some (e.g. [15], [18]) have particularly dealt with applicability of the synthesis techniques on mobile devices, some others (e.g. [12], [19], [22], [25], [26]) have concentrated on the duration modeling, whereas some (e.g. [14], [28]) have focused on achieving prosody in the synthesized speech. On the other hand, the number of studies focusing on pronunciation disambiguation is very limited. In this subject, due to their approach of identifying and handling the examples, [24] and [32] can be considered as biblical resources. Moreover, they provide almost a complete set of interesting cases for ambiguities in Turkish pronunciation. In [20], a statistical approach for pronunciation disambiguation was proposed. In [31] and [36], additional exceptional cases (i.e. cases for which pronunciation ambiguity occurs) in Turkish were discussed; a practical framework for pronunciation ambiguity resolution was proposed.

The main motivation of this study can be summarized as follows: For Turkish text-to-speech synthesis applications, the need for the creation of a pronunciation lexicon together with a rule set is unarguable. As long as this lexicon is enriched, the pronunciation accuracy of a text-to-speech synthesizer depending on this infrastructure (i.e. the lexicon and the rule set) would get better. The first step in achieving a robust infrastructure would be to identify the problematic/exceptional cases, which have already been done in [31] and [36]. In this study, we carry out the next step, which is nothing but the determination of the correct order of lexicon enlargement for rapid coverage (i.e. the ideal order of inclusion of exceptional syllables in order to achieve maximum pronunciation ambiguity resolution capability with minimum effort). In other words, our aim in this study is to identify which cases are encountered most frequently in daily used Turkish language. To our belief, the results of this study will serve as a guideline for following research studies about the prioritization of lexicon enrichment.

The outline of this paper is as follows: After this brief introduction section, in Section 2, we will try to revisit the main cases where pronunciation ambiguity occurs in Turkish; and at the same time classify them. In Section 3, we will give the results of the statistical analysis of the occurrence rates of the identified 7 main groups. In Section 4, we will increase the depth of this statistical analysis by considering the so-far-identified 154 syllables distinctly. Section 5 will include comments and discussions about the analysis results together with potential future work.

2 Pronunciation Ambiguity in Turkish

2.1 Historical Background

Even though it is claimed that "the current Turkish alphabet is phonetic" (i.e. the grapheme-to-phoneme mapping is one-to-one), especially for the words imported from foreign languages, such as Arabic, Persian and French, many occurrences of one-to-many grapheme-to-phoneme mappings can be found [36]. Certainly, the complexity of the grapheme-to-phoneme mapping is not as dramatic as in French or in English (e.g. there exist many unpredictable pronunciations in these languages such as the pronunciation of the 4-gram "ough" in the words "rough", "cough", "dough", "tough", "though", "through", "thorough"). Moreover, as demonstrated in [31] and [36], it is possible to handle almost all exceptional cases in Turkish by means of accent signs, which are introduced on vowels. On the other hand, regardless of its complexity, it is unarguable that the occurrence of pronunciation ambiguities constitutes a considerable ratio.

Regarding the phonemes in modern standard Turkish, there have been several studies [37]-[39] which have been performed by experts of linguistics. All these studies agree on the fact that the number of phonemes is much more than the number of symbols in the current Turkish alphabet. In one of the most respected studies on this subject [39], 44 phonemes have been identified. On the other hand, the current Turkish alphabet, which is based on the Latin alphabet, consists of 29 letters. During the adoption of the Latin alphabet in 1928 (the so-called "Alphabet Revolution"), though there were proposals of 32-letter alphabets, a set of 29 letters was considered to be sufficient [40].

In addition to 29 letters, an accent sign (i.e. "^") was considered to be necessary and sufficient. This sign used to have multiple purposes: increasing the duration of the current vowel in some occasions (as in the word "bâriz (obvious)", for which the duration of the letter a is longer than normal), palatalization of the preceding consonant (as in the word "kâğıt (paper)", for which the letter k is palatalized), or both (as in the word "kâbus (nightmare)", for which the letter k is palatalized and the duration of the letter a is longer than normal). Presently, this accent sign has become almost obsolete in practice due to two main factors: (i) untruthful rumors that the usage of this accent sign was cancelled by the Turkish Language Council in 1980s, (ii) for written communication, the wide-spread usage of media (such as e-mail, SMS, etc.) which did not support the accent sign.

For human readers, who perform pattern recognition and resolve pronunciation ambiguities automatically (and unconsciously), the pronunciation ambiguities do not constitute a problem in Turkish as in some other languages. On the other hand, when the speech is synthesized by machinery, the introduction of some mechanisms (for the machinery to identify these ambiguities) becomes compulsory. Otherwise, the quality of the synthesized speech would be irritating for the listeners; and it might even yield lexical and/or syntactical misunderstandings in some cases.

2.2 Cases of Pronunciation Ambiguity

As stated before, in most of the studies conducted so far, pronunciation ambiguity in Turkish has not been handled, or not even mentioned. For example in [29], the authors claimed to obtain reasonable synthesized results. On the other hand, since they did not mention pronunciation ambiguity in Turkish, how they achieved what they claimed is a big question mark.

In [31] and [36], exceptional syllables (i.e. the syllables for which the grapheme-to-phoneme conversion mapping is one-to-many) have been identified as follows (Throughout the following items, the example words are given in syllabified form in order to provide better understanding, especially to the non-Turkish speaking readers):

1) Syllables ending with the letters a, e, i, o, u, ü: In such syllables, the relevant letter might be pronounced normally (e.g. as in the words a-tak (attack), e-tek (skirt), i-nek (cow), o-to-büs (bus), u-fuk (horizon), ü-mit (hope)); or in lengthened form (e.g. as in the words a-şık (lover, folk poet), me-mur (government officer), i-kaz (warning), li-mo-ni (lemonish), u-di (lute player), mü-min (religious person, believer)).

2) Syllables ending with the digrams al, ol, ul: In such syllables, the letter l might be pronounced velar (e.g. as in the words al-kış (handclap), bol (numerous, copious), dul (widow)); or alveolar (e.g. as in the words al-kol (alcohol), gol (goal), ma-kul (reasonable)).

3) Syllables starting with the digrams la, lo, lu: In such syllables, the letter l might be pronounced velar (e.g. as in the words la-la (life-coach of the Ottoman Prince), ba-lo (party, ball), o-luk (groove)); or alveolar (e.g. as in the words lam-ba (lamp), fi-lo (fleet), bil-lur (crystal)).

4) Syllables starting with the letters k, g: In such syllables, the relevant letter might be pronounced velar (e.g. as in the words kar-tal (eagle), ga-ga (beak); or palatal (e.g. as in the words ka-ğıt (paper), ga-vur (giaour)).

5) Syllables ending with the digram at: In such syllables, the digram at might be pronounced normally (e.g. as in the words kat (floor, flat), yat (yacht)); or softly as if there is the phoneme e in between (i.e. similar to the aet triphone but in a rapid manner) (e.g. as in the words sa-at (clock), sıh-hat (health)).

6) Syllables starting with the digram na: In such syllables, the digram na might be pronounced normally (e.g. as in the words nar (pomegranate), naz (whims)); or softly as if there is the phoneme e in between (i.e. similar to the nea triphone but in a rapid manner) (e.g. as in the word ma-na (meaning)).

7) Syllables ending with the digram el, em, en: In such syllables, the letter e might be pronounced normally (e.g. as in the words bel-li (definite), em-zik (pacifier), en-gin (profound)); or widely (e.g. as in the words bel-ge (document), ma-tem (mourning), mü-ren (muraena)).

As described above, for these exceptional syllables, generally there exist two different pronunciations. On the other hand, it should be noted that some syllables might fall into more than one category according to the classification given above. For example, the syllable ka belongs to the 1st and the 4th classes at the same time. As a result of this, there exist four different pronunciations of this syllable for different occasions:

1) kaba (rough); for which the letter k is pronounced velar, and the letter a is pronounced normally.

2) kabiliyet (capability); for which the letter k is pronounced velar, and the letter a is pronounced in lengthened form.

3) kağıt (paper); for which the letter k is pronounced palatal, and the letter a is pronounced normally.

4) katip (clerk); for which the letter k is pronounced palatal, and the letter a is pronounced in the lengthened form.

2.3 Proposed Method and Architecture for Pronunciation Ambiguity Detection/Resolution

As stated and demonstrated via numerous examples in [32], for complete and accurate pronunciation ambiguity resolution in Turkish, it is compulsory to perform syntactical analysis in addition to lexical analysis (e.g. for some miscellaneous cases such as the pronunciation ambiguity resolution of isographic words; for example the word sol (left), for which the letter l is pronounced velar; and the word sol (musical note G), for which the letter l is pronounced alveolar).

On the other hand, since Turkish is an agglutinative language, syntactical analysis is a very complicated task. Due to this fact, in [36], a practical method for pronunciation ambiguity resolution (without rigorous syntactical analysis) has been proposed. Certainly, this method would not be able to perform the resolution of some miscellaneous cases such as the isographic words; but it is able to resolve the problems listed in Section 2.2 (such as the identification of the syllable bal in the word bal (honey), for which the letter l is pronounced velar, and in the word istikbal (future), for which the letter l is pronounced alveolar), which constitute the majority of the pronunciation ambiguity problems in Turkish. Moreover, it should be noted that for some cases, syntactical analysis by itself would not be sufficient; more advanced and intelligent methods for contextual identification might be. (E.g. for the resolution of the statement “Karlı bir yıl geçirdik [We experienced a very profitable/snowy year]”, the context of the overall text shall be identified. If it is a text about the meteorological information, the word kar shall be identified as kar (snow), for which the letter k is pronounced normally; if it is about finance, then the word kar shall be identified as kâr (profit), for which the letter k is palatalized.)

In [36], the symbology seen in Table 1 was proposed in order to achieve a phonetic representation. Examples regarding the usage of this phonetic representation are listed in Table 2.

Table 1

Proposed additional symbols and their definitions (according to [36])

Normal	Letter pronounced normally	Aa	Ee	İi	Oo	Uu	Üü
Long	Letter pronounced in a longer manner	Ââ	Êê	Îî	Ôô	Ûû	ßÿ
Thin	Inside a syllable: - The a, o, and u letters succeeding the alveolar l letter; - The a, o, and u letters succeeding the palatal k or g letters; - The a letter included in the na diphone, which is pronounced as the nea triphone.	Áá	–	–	Óó	Úú	–
Long and Thin	The a and u letters satisfying the conditions of being “long” and “thin” simultaneously.	Ãã	–	–	–	Ýý	–
Soft	Inside a syllable: - The a, o, and u letters preceeding the alveolar l letter; - The a letter included in the at diphone, which is pronounced as the aet triphone.	Àà	–	–	Òò	Ùù	–
Wide	The widely pronounced e letter.	–	Ëë	–	–	–	–

Table 2

Examples regarding the usage of the proposed symbols in [36]

	a	**e**	**i**	**o**	**U**	**ü**
Normal	araba (car)	etek (skirt)	inek (cow)	otomobil (automobile)	uzun (long)	ütü (iron)
Long	âşık (lover)	têmin (obtain)	îkaz (warning)	limônî (lemonish)	ûdî (lute player)	mÿmin (believer)
Thin	láma (llama), káğıt (paper), gávur (giaour)	–	–	lómboz (porthole)	billúr (crystal), sükút (silence)	–

Long and Thin	lãle (tulip), kãbus (nightmare), yegãne (unique), mânã (meaning)	–	–	–	ulýfe (salary of the soldiers in the Ottoman Empire), sükýnet (silence)	–
Soft	ihmàl (ignorance), itaàt (obey)	–	–	gòl (goal)	kabùl (acceptance)	–
Wide	–	dirhëm (drachmai)	–	–	–	–

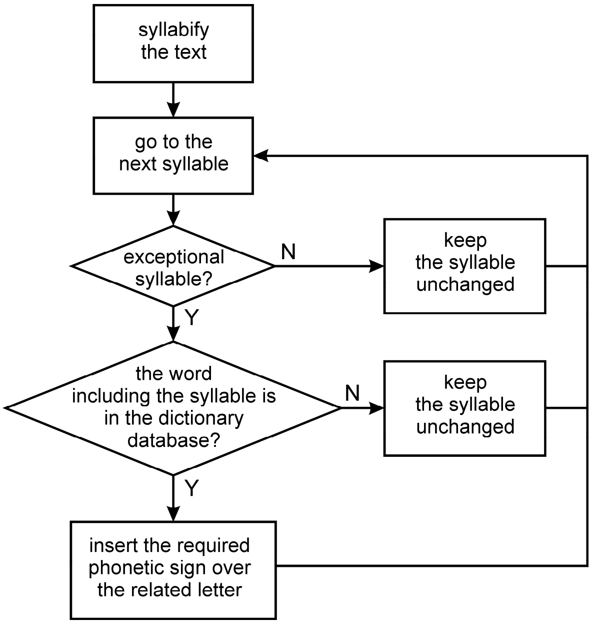
syllabify the text
go to the next syllable
exceptional syllable?
N
keep the syllable unchanged
Y
the word including the syllable is in the dictionary database?
N
keep the syllable unchanged
Y
insert the required phonetic sign over the related letter

Figure 1
Flowchart of the proposed pronunciation ambiguity detection/resolution methodology

The proposed method for pronunciation ambiguity resolution is straightforward, as seen in Fig. 1. The prerequisite for complete/correct performance of this method is the existence of a lexicon identifying the pronunciation of the words including exceptional syllables. The algorithm syllabifies the text to be synthesized. One by one, it controls whether each encountered syllable is exceptional or not. If a syllable is exceptional, and if the word containing that syllable is inside the lexicon (in case that such a lexicon is constructed), then the pronunciation of the relevant syllable is identified to be exceptional. The very basic structure of such a lexicon is given in Table 3.

Table 3
The structure of the pronunciation lexicon and some examples

Exceptional Word	Number of Exceptional Syllables	Exceptional Syllable Position(s)	Pronunciation(s) in Relevant Syllable(s)
`arazi` (field)	1	{2}	{1}
`makul` (reasonable)	2	{1,2}	{1,3}
`samimi` (sincere)	1	{2}	{1}

The fields of such a lexicon can be explained as follows: Each row of the lexicon contains a separate word; the number of exceptional syllables in that word, and the positions of these syllables. The pronunciations of such syllables are coded by means of an enumerated type (e.g. 1 standing for the lengthening of the vowel, 2 standing for palatalization of the consonant at the beginning, 3 standing for alveolarization of the consonant at the end, etc.). By means of such a structure, it is possible to model the words containing more than one exceptional syllable (such as `makul` (reasonable), represented and pronounced as `mâkùl`); or the words containing a unique syllable more than once, whose occurrences are pronounced differently (such as `samimi` (sincere), represented and pronounced as `samîmi`; or `hakiki` (real, original), represented and pronounced as `hakîki`).

At this point, it should be noted that even though the pronunciation check/control activity is based on syllabification and syllables, the framework does not imply that the speech synthesis shall be concatenative and syllable based. In other words, the proposed method can be integrated with any speech synthesizing technique.

Another remark is the possibility of extension of this lexicon by introducing new columns, such as the positioning of the intonation and stress for prosody in speech synthesis.

3 First Level Statistical Analysis and Results

As stated in [41], language statistics have a quite important role in speech synthesis and recognition applications for high fidelity. In this chapter, we try to give figures of merit about how frequently the exceptional syllables occur in the Turkish of daily life. For this purpose, we have parsed 48 books (short stories, novels, essays and scenarios written by several amateur and professional writers) including a total of 1,529,647 words.

As the basis of the statistical analysis in this study, we implemented a so-called “syllable hunter” script in MATLAB, which depends on the syllabification algorithm defined in [36]. The main idea of this algorithm is based on determining the locations of the vowels through the words, since each Turkish syllable contains one vowel. The algorithm also handles the syllabification of some imported compound words, which linguistically have Latin origins (e.g. `elektronik` (electronics) to be syllabified correctly as `e-lek-tro-nik` but not as `e-lekt-ro-nik`). Our "syllable hunter" gets each word one by one from the parsed source text and extracts the syllables into a syllable pool in accordance with the flowchart given in Fig. 2.

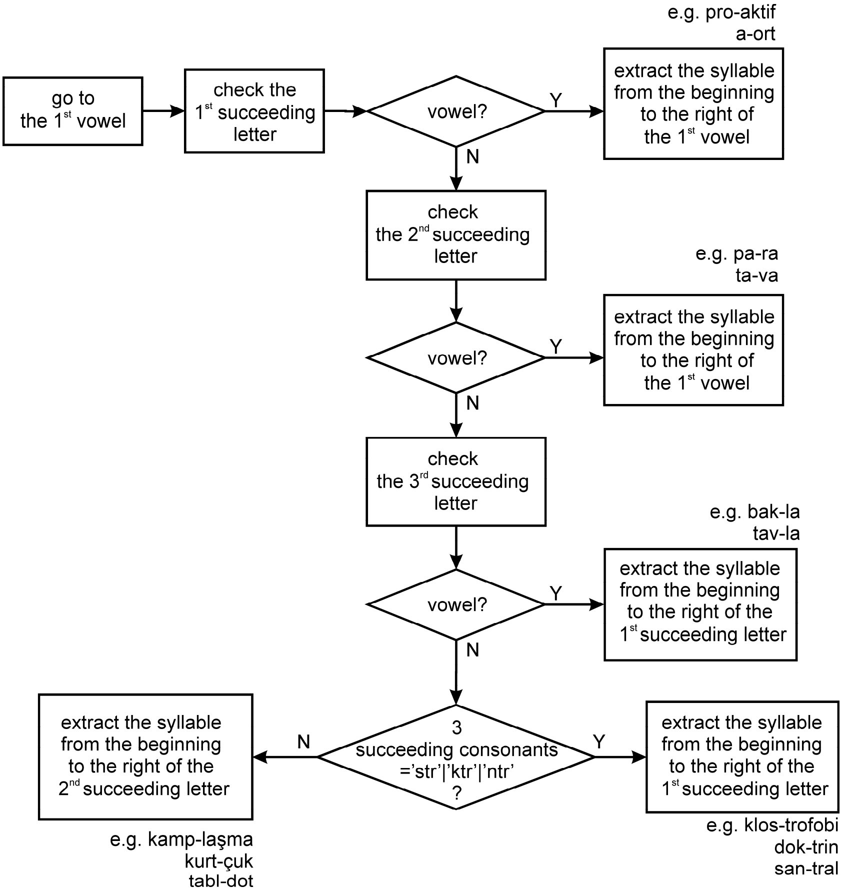
go to the 1st vowel
check the 1st succeeding letter
vowel?
Y
N
extract the syllable from the beginning to the right of the 1st vowel
e.g. pro-aktif a-ort
check the 2nd succeeding letter
vowel?
Y
N
extract the syllable from the beginning to the right of the 1st vowel
e.g. pa-ra ta-va
check the 3rd succeeding letter
vowel?
Y
N
extract the syllable from the beginning to the right of the 1st succeeding letter
e.g. bak-la tav-la
3 succeeding consonants ='str'|'ktr'|'ntr' ?
N
Y
extract the syllable from the beginning to the right of the 2nd succeeding letter
e.g. kamp-laşma kurt-çuk tabl-dot
extract the syllable from the beginning to the right of the 1st succeeding letter
e.g. klos-trofobi dok-trin san-tral

Figure 2
Flowchart of the syllabification algorithm for Turkish [36]

By means of the "syllable hunter", we syllabified the entire word-set, and obtained all the distinct syllables together with their numbers of occurrences in the processed texts. We found that the aforementioned 48 books contain 4,043,954 syllables in total. Next, we analyzed the syllables in order to obtain the statistics of the exceptional syllables, where the exceptional syllables were identified according to the rules given in Section 2.2.

As a second step, we classified the syllables into four different groups according to their lengths. In Turkish, a syllable might consist of at least 1 letter, and at most 4 letters. In recent years, some words with 5-letter syllables (e.g. `tvist` (twist), `frenk` [French or more generally European, Western], etc.) have been imported and adopted. But since the occurrence rate of the 5-letter syllables is relatively small, we have not considered them in this study.

The charts in Fig. 3 depict the overall syllable distribution statistics of the processed texts in this study, comparatively with [29]. Except the 5-letter syllables (which have been ignored by us), it can be seen that our results are in almost perfect agreement with [29]. This means that our data constitutes a sufficiently-large set, over which confident statistical analyses can be performed and meaningful results can be obtained.

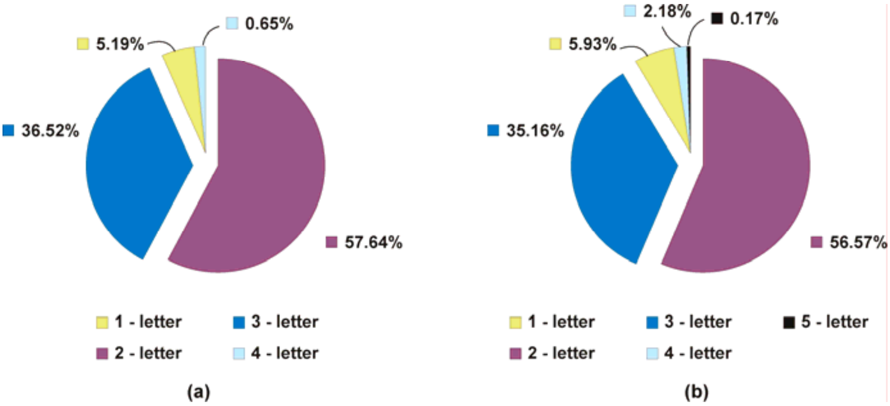

Figure 3

Overall syllable distribution statistics: results of this study (a) vs. [29] (b)

Table 4 shows how frequently appear the syllables ending with the letters '`a`, `e`, `i`, `o`, `u`, `ü`' as separately and as a whole. Due to phonological features of Turkish, syllables ending with vowels constitute the majority. Hence, it is not surprising for us that more than 51% of all possible syllables end with these 6 vowels (As a matter of fact, the remaining 2 vowels `ı` and `ö` are not as frequent as `a` or `e`). On the other hand, it is apparent from Fig. 3 that about 57% of syllables are 2-letter. Thus, having the 2-letter syllable dominance in this class is expected. Moreover, the majority of the 2-letter syllables belonging to this class are the ones ending with `a`, `e` and `i`. These will be investigated in detail in the upcoming Sections.

Table 4

Frequencies (%) of the syllables ending wıth the letters 'a, e, i, o, u, ü'

	1-letter	2-letter	3-letter	4-letter	Total
'a'	1.4813	15.7036	0.0214	0.0003	17.2066
'e'	0.6360	12.3194	0.0149	0.0002	12.9705

'i'	1.1519	10.0035	0.0170	0.0002	11.1726
'o'	1.1845	1.5677	0.0203	0.0001	2.7725
'u'	0.3161	4.2767	0.0048	0.0000	4.5975
'ü'	0.1324	2.1939	0.0004	0.0000	2.3267
				Group:	51.0465

Table 5 exhibits the statistics for the syllables ending with the digrams ‘`al`, `ol`, `ul`’. Similarly, Table 6 lists the frequencies for the syllables starting with the digrams ‘`la`, `lo`, `lu`’. Table 7 gives the frequencies of the syllables starting with the letters ‘`k`, `g`’. Table 8 shows the frequencies of the syllables ending with the digram ‘`at`’. Table 9 presents the frequencies of the syllables starting with the digram ‘`na`’. Table 10 shows the frequencies of the syllables ending with the digrams ‘`el`, `em`, `en`’.

Table 5

Frequencies(%) of the syllables ending wıth the digrams ‘al, ol, ul’

	1-letter	2-letter	3-letter	4-letter	Total
'al'	-	0.2504	0.5063	0.0023	0.7590
'ol'	-	0.4921	0.1486	0.0031	0.6439
'ul'	-	0.0009	0.2417	0.0000	0.2426
				Group:	1.6455

Table 6

Frequencies(%) of the syllables starting wıth the dıgrams ‘la, lo, lu’

	1 letter	2 letters	3 letters	4 letter	Total
'la'	-	2.3886	1.4385	0.0038	3.8310
'lo'	-	0.0433	0.0228	0.0018	0.0679
'lu'	-	0.3219	0.2725	0.0001	0.5945
				Group:	4.4934

Table 7

Frequencies(%) of the syllables starting wıth the letters ‘k, g’

	1-letter	2-letter	3-letter	4-letter	Total
'k'	-	3.7566	2.5156	0.1018	6.3739
'g'	-	1.8090	1.1707	0.0605	3.0403
				Group:	9.4142

Table 8

Frequencies (%) of the syllables ending wıth the dıgram ‘at’

	1-letter	2-letter	3-letter	4-letter	Total
'at'	-	0.1024	0.3880	0.0004	0.4908

Table 9

Frequencies (%) of the syllables starting with the digram ‘na’

	1-letter	2-letter	3-letter	4-letter	Total
'na'	-	0.9818	0.1906	0.0041	1.1765

Table 10

Frequencies (%) of the syllables ending with the digram ‘el, em, en’

	1-letter	2-letter	3-letter	4-letter	Total
'el'	-	0.0832	0.4198	0.0047	0.5078
'em'	-	0.0224	0.2534	0.0008	0.2765
'en'	-	0.1005	1.9986	0.0085	2.1076
				Group:	2.8919

General observations about these statistics can be summarized as follows:

(i) As stated above, for pronunciation disambiguation, special attention shall be devoted to the 2-letter syllables ending with vowels, particularly the ones ending with a, e and i.

(ii) Since the 1-letter syllables have to be vowels, Tables 5 to 10 have zero entries for 1-letter column as expected. As seen in Fig. 3, 1-letter syllables constitute almost 6% of the whole set. Since we have 6 of 8 vowels in Table 4, we can conclude that 1-letter syllables belonging to this group also require special attention.

(iii) It is very rare that a 3- or 4-letter syllable ends with a vowel; which can also be observed from Table 4. Hence, such syllables might have small importance.

(iv) As seen from Table 6, 2- and 3-letter syllables starting with the digram la has considerable frequency.

(v) As seen from Table 7, 2- and 3-letter syllables starting with the letter k has considerable frequency. Such syllables starting with the letter g are also of importance.

4 Second Level Statistical Analysis and Results

In [36], it has been identified that there exist at least 154 exceptional syllables which cause pronunciation ambiguity in Turkish. In this chapter, we focus our attention to these syllables, and give the statistical results for the frequencies of these 154 exceptional syllables. Table 11 lists the frequencies of these syllables (sorted from the most frequent to the least). It can be seen that syllables ending with a and i dominate the top positions of the list. It can be seen that some syllables belonging to more than one class (i.e. the classes mentioned in Section 2.2); such as la, ka and na have a considerable occurrence rate.

Table 11
Frequencies (%) of the 154 exceptional syllables (sorted from the most frequent to the least)

syllable	frequency	syllable	frequency	syllable	frequency
la	2.3886	kal	0.1416	pen	0.0220
di	1.6241	ga	0.1336	dol	0.0213
da	1.6144	mü	0.1186	sem	0.0182
ya	1.5514	vi	0.1064	bol	0.0161
ka	1.4985	cu	0.1063	gar	0.0160
a	1.4813	ren	0.1055	kam	0.0157
ma	1.3836	sen	0.1034	tel	0.0156
ra	1.1818	at	0.1024	tem	0.0154
i	1.1519	kan	0.0963	kut	0.0152
ri	1.0545	şu	0.0958	kun	0.0140
na	0.9818	hi	0.0933	tal	0.0140
bi	0.9293	men	0.0894	dem	0.0136
ni	0.8790	kat	0.0825	lon	0.0128
ki	0.8440	fi	0.0788	ral	0.0126
me	0.8109	ber	0.0785	sol	0.0118
ba	0.8035	bul	0.0713	bal	0.0109
du	0.7943	laş	0.0689	gan	0.0108
lar	0.7902	lur	0.0676	kum	0.0101
li	0.7885	yen	0.0659	cen	0.0100
ta	0.7688	yal	0.0612	rem	0.0090
bu	0.7615	hal	0.0611	şal	0.0081
ha	0.7507	kul	0.0605	kem	0.0076
sa	0.6981	bel	0.0597	nal	0.0062
si	0.6532	lun	0.0575	pal	0.0062
ti	0.6227	luk	0.0556	rol	0.0058
te	0.6201	lat	0.0547	lut	0.0055
den	0.5437	sal	0.0542	kel	0.0047
mi	0.4742	nem	0.0528	las	0.0046
nu	0.4026	hat	0.0524	cer	0.0043
ca	0.3924	hem	0.0513	yem	0.0042
ken	0.3627	vu	0.0446	yel	0.0042
lan	0.3341	lo	0.0433	cel	0.0039
lu	0.3219	lah	0.0431	laç	0.0038
u	0.3161	hu	0.0428	gal	0.0033
za	0.3150	fen	0.0424	gul	0.0032
şa	0.2869	kah	0.0423	zal	0.0032

syllable	frequency	syllable	frequency	syllable	frequency
tu	0.2804	lak	0.0421	zem	0.0026
ku	0.2681	zu	0.0392	fel	0.0024
ça	0.2596	lam	0.0391	fal	0.0023
al	0.2504	mem	0.0385	çem	0.0021
kar	0.2488	pi	0.0379	cem	0.0021
ru	0.2429	lay	0.0378	bem	0.0020
pa	0.2215	zen	0.0366	pul	0.0019
su	0.1914	lum	0.0341	ul	0.0009
va	0.1869	mal	0.0338	gat	0.0008
ben	0.1858	dal	0.0307	fol	0.0008
tü	0.1755	mo	0.0288	tol	0.0007
mu	0.1733	kol	0.0266	lom	0.0005
fa	0.1701	val	0.0265	pol	0.0004
ci	0.1561	lup	0.0239	ja	0.0000
zi	0.1497	sul	0.0236		
ten	0.1431	rat	0.0230		

We performed another analysis in order to identify the coverage rate. In other words, we tried to identify how much pronunciation disambiguation capability would be achieved by adding the words with the most occurring syllables to the pronunciation lexicon. Here is what we obtained: The most occurring 12 syllables constitute 50% of occurrences of whole exceptional syllables; similarly 50 of them constitute 90%, and 100 of them constitute 99% of exceptional occurrences. This trend is illustrated in more detail in Fig. 4.

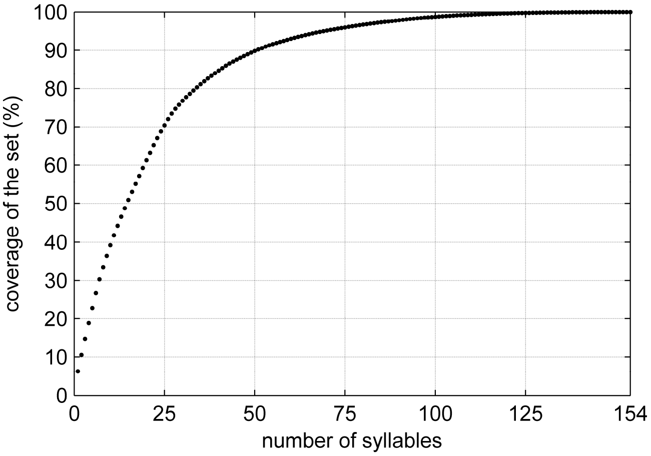

Figure 4

Percent ambiguity resolution coverage curve

We can rephrase the findings of this Section as follows: If a researcher wants to enrich his/her lexicon as defined in this study (Section 2.3), then he/she must start adding all words including the syllable `la`; and continue this process according to the order given in Table 11. The addition of all words having the first 50 syllables would give 90% pronunciation disambiguation capability, which seems to yield a more or less optimum efficiency (maximum coverage with minimal lexicon enrichment effort). The inclusion of the 100 most frequent exceptional syllables would imply 99% coverage, which means that the last 54 entries of Table 11 might be neglected practically.

Conclusions and Future Work

In this study, we have tried to identify the exceptional syllables for which the grapheme-to-phoneme mapping is not one-to-one, as well as the occurrence rates of these syllables. On the other hand, it should be noted that the given statistics refer to the total occurrences of these exceptional syllables (i.e. both the normal/default pronunciation and the abnormal/extraordinary pronunciation cases are counted together). For more granularity about the rate of extraordinary pronunciations, additional analyses are required; and for these analyses, the aforementioned pronunciation lexicon should be complete. Our near- and mid-term plans are to enrich the lexicon for the most occurring syllables, and try to come up with more statistics about such syllables (i.e. the rate of extraordinary pronunciation for these syllables).

In Table 12, the effectiveness of the proposed technique is demonstrated by means of some example sentences for which the pronunciation ambiguity for the existing exceptional syllables are resolved and the phonetic representations are obtained.

Table 12

Example sentences demonstrating the effectiveness of the proposed technique

Sample Text	Relevant Proposed Phonetical Representation
Tesislerden yararlanan tüm memurların, seyahatleri ve tatilleri esnasında bu hususa dikkat etmeleri gerektiği açıklandı.	Têsislerden yararlanan tüm mêmurların, seyahàtleri ve tâtilleri esnâsında bu husûsa dikkàt etmeleri gerektiği açıklandı.
Afet bölgesini beraberindeki heyetle ziyaret eden Hakkari Valisi, kabul ettiği felaketzedelere bugüne kadar sükunet ve fedakarlıkla göğüs gerdikleri problemlerin derhal giderileceğini, bu konuda hiç bir ihmalkarlığa tahammül edilmeyeceğini bildirdi.	Âfet bölgesini berâberindeki heyetle ziyâret eden Hakkãri Vâlisi, kabùl ettiği felãketzedelere bugüne kadar sükýnet ve fedâkárlıkla göğüs gerdikleri problemlerin derhàl giderileceğini, bu konuda hiç bir ihmàlkárlığa tahammül edilmeyeceğini bildirdi.
Zamanında belediyeye bağlı Zabıta Amirliği tarafından düzenlenmekte olan mahalli lale festivalinin, bu yıl Yeşil Vadi olarak da bilinen bölgede Kağıt Fabrikası'nın karşısındaki alanda valilik tarafından düzenleneceği bildirildi.	Zamânında belediyeye bağlı Zâbıta Âmirliği tarafından düzenlenmekte olan mahàllî lãle festivàlinin, bu yıl Yeşil Vâdi olarak da bilinen bölgede Káğıt Fabrikası'nın karşısındaki alanda vâlilik tarafından düzenleneceği bildirildi.

Tüm sözlü ikazlara ve yazılı belgelere rağmen, Nisan-Haziran döneminde İran sınırı üzerinden gerçekleşen anormal mülteci akımına karşı acil bir önlem alınmadı.	Tüm sözlü îkazlara ve yazılı bëlgelere rağmen, Nîsan-Hazîran döneminde Îran sınırı üzerinden gerçekleşen anormàl mültecî akımına karşı âcil bir önlem alınmadı.
Cesaretleri ile nam salmış olan Cezayir korsanları, rutubetten kaynaklı suhulet düşüklüğü nedeniyle, kalyonlarının seyrini normalden daha düşük süratle, narin ve nazik bir şekilde idame ettiriyorlardı.	Cesâretleri ile nam salmış olan Cezâyir korsanları, rutûbetten kaynaklı suhûlet düşüklüğü nedeniyle, kàlyonlarının seyrini normàlden daha düşük süràtle, nârin ve nâzik bir şekilde idâme ettiriyorlardı.

The results showed that among the exceptional syllables, especially for the 1-letter and 2-letter syllables ending with the letters `a`, `e` and `i,` are the most frequent ones generally. At this point, we make the following remarks based on our personal experiences: Even though the syllables ending with `e` are very frequent, the phenomenon of lengthening the vowel `e` is very rare. In other words, there are only a limited number of words (in the order of a couple) for which the vowel `e` is pronounced in lengthened form (such as `memur` (government officer), represented and pronounced as `mêmur`; `tesis` (facility), represented and pronounced as `têsis`; `temin` (obtainment), represented and pronounced as `têmin`). Hence, for the syllables ending with the letter `e`, it is very easy to complete the pronunciation lexicon. On the other hand, there are numerous words for which the vowel `a` is pronounced in lengthened form (in the order of thousands) and for which the vowel `i` is pronounced in lengthened form (in the order of hundreds). Hence, it will be a time- and effort-consuming task to identify all such words and include them in the lexicon. In addition, due to their being elements of multiple classes, syllables `la` and `ka` (and the words including them) are very frequent, and they also require attention.

An important point to be emphasized is that the proposed method is not able to resolve ambiguities despite its ability to detect them for the homeomorphic/isographic words (e.g. `kar` (snow/profit), `ama` (but/blind), `adet` (number/habit), etc.). As stated earlier in Section 2.3, syntactic analysis (moreover, in some instances, even contextual meaning analysis) is required for the resolution of ambiguities caused by the homeomorphic/isographic words. On the other hand, another analysis is also performed in order to have a qualitative idea about the occurrence rate of such words in meaningful Turkish texts. As seen in Table 13, frequencies of such words are computed as negligible for a test performed by using a text of 1,549,647 words. Hence, it can be concluded that the coverage of the proposed technique is quite good considering its practicality.

At this point, the following remark shall be made in order to prevent any misinterpretations of the results given in Table 13. The numbers given in Table 13 indicate the number of words starting with the relevant pattern. For example, the number 12,752 for the pattern `kar` means that 12,752 words starting with the syllable `kar` were encountered in the text; accounting not only the isolated homemomorphic word `kar` (snow or profit) but also the words such as `karşı` (against), karşıt (opposite), kartal (eagle), karton (cartoon), karmaşık (complicated), etc. together with their all suffixed forms. The proposed method already resolves all the pronunciation ambiguities for the words karşı, karşıt, kartal, karmaşık, etc. and all their suffixed forms; but only gets stuck for the occurences of kar and its suffixed forms (which is only a very limited percent of the number 12,752). For the occurrences of kar and its suffixed forms, the proposed methods leave them as is (i.e. all the occurrences are to be pronounced as if the word means snow); hence, the occurrences of kar with the meaning profit will be misrepresented and mispronounced, and certainly these constitute a much lower percentage of the number 12,752. The same arguments are valid also for the other homeomorphic words seen in Table 13. Considering this, the percentage of misrepresentations and mispronunciations with the proposed method are quite low (i.e. the total number seen in Table 13 is a very exaggerated upper bound; the number of the exact misrepresentations and mispronunciations would probably be much less than 1/10 of the total number given in Table 13).

Table 13
Frequencies of the homeomorphic/isographic words

Word	Pronunciation and Meaning	Pronunciation and Meaning	Occurance
adet	adet (number)	âdet (habit)	355
ala	ala (colorful)	âlã (superb)	2619
ali	ali (a proper name)	âlî (lofty)	936
ama	ama (but)	âmâ (blind)	7504
aşık	aşık (compete)	âşık (lover)	411
atıl	atıl (pounce)	âtıl (idle)	424
dahi	dahî (even)	dâhi (genius)	393
hala	hala (aunt)	hâlã (still)	959
kar	kar (snow)	kár (profit)	12752
mal	mal (goods)	màl (cost)	908
sol	sol (left)	sòl (note G)	1032
usul	usul (quitely)	usùl (method)	197
varis	varis (varicosis)	vâris (inheritor)	6
		Total	28496

To our belief, the results of this study might additionally serve as a guideline for researches related with different topics:

(i) General syllable statistics might find application areas such as statistical ambiguity resolution in optical character recognition, or even in speech recognition.

(ii) These statistics might also be considered for the computation of syllable-based entropy calculation of the Turkish language. Such an entropy value might be used in information theoretical research studies.

(iii) The syllables, their frequencies and their lengths might also provide input for the definition of new readability metrics of Turkish texts.

Moreover, even though the statistical data provided here are focused in Turkish, our approach might also be applied to another language in future studies for similar purposes.

References

[1] Ozum, Y.: A Speech Synthesis System for Turkish Language Based on the Concatenation of Phonemes taken from Speaker. M.Sc. Thesis, Middle East Technical University, Ankara, Turkey, 1993

[2] Erer, M. S.: Text-to-Speech in Turkish Language by Using a Mixed Speech Synthesis Method. M.Sc. Thesis, Istanbul Technical University, Istanbul, Turkey (in Turkish), 1994

[3] Güven, K.: PC Based Speech Synthesis for Turkish. M.Sc. Thesis, Çukurova University, Adana, Turkey, 1994

[4] Öztaner, S. M.: A Word Grammar of Turkish with Morphophonemic Rules. MSc Thesis, Middle East Technical University, Ankara, Turkey, 1996

[5] Ayhan, K.: Text-to-Speech Synthesizer in Turkish Using Non Parametric Techniques. M.Sc. Thesis, Middle East Technical University, Ankara, Turkey, 1998

[6] Salor, Ö.: Signal Processing Aspect of Text to Speech Synthesizer in Turkish. MSc Thesis, Middle East Technical University, Ankara, Turkey, 1999

[7] Bozkurt, B.: Reading Aid for Visually Impaired (A Turkish Text-to-Speech System Development) MSc Thesis, Bogazici University, Istanbul, Turkey, 2000

[8] Abdullahbeşe, E.: Fundamental Frequency Contour Synthesis for Turkish Text-to-Speech. MSc Thesis, Bogazici University, Istanbul, Turkey, 2001

[9] Ömür, Ç.: Concatenative Speech Synthesis Based on a Sinusoidal Speech Model. MSc Thesis, Middle East Technical University, Ankara, Turkey, 2001

[10] Eker, B.: Turkish Text to Speech System, M.Sc. Thesis, Bilkent University. Ankara, Turkey, 2002

[11] Özen, Ş. S.: Turkish Text to Speech Synthesis. MSc Thesis, Hacettepe University, Ankara, Turkey (in Turkish) 2002

[12] Şayli, Ö.: Duration Analysis and Modelling for Turkish Text-To-Speech Synthesis. MSc Thesis, Bogazici University, Istanbul, Turkey, 2002

[13] Oskay, B.: Automatic Modeling of Turkish Prosody. MSc Thesis, Middle East Technical University, Ankara, Turkey, 2002

[14] Vural, E.: A Prosodic Turkish Text-to-Speech Synthesizer. MSc Thesis, Sabanci University, Istanbul, Turkey, 2003

[15] Aktan, O.: A Single Chip Solution for Text-to-Speech Synthesis. MSc Thesis, Bogazici University, Istanbul, Turkey, 2004

[16] Sak, H.: A Corpus-based Concenative Speech Synthesis System for Turkish. MSc Thesis, Bogazici University, Istanbul, Turkey, 2004

[17] Karlı, A.: A Turkish Text-to-Speech Synthesizer for a Set of Sentences. MSc Thesis, Ankara University, Ankara, Turkey (in Turkish) 2005

[18] Ünaldı, İ.: Turkish Text to Speech Synthesis System for Mobile Devices. MSc Thesis, Hacettepe University, Ankara, Turkey (in Turkish) 2007

[19] Öztürk, Ö.: Modeling Phoneme Durations and Fundamental Frequency Contours in Turkish Speech. PhD Dissertation, Middle East Technical University, Ankara, Turkey, 2005

[20] Külekçi, M. O.: Statistical Morphological Disambiguation with Application to Disambiguation of Pronunciations in Turkish, PhD Dissertation, Sabancı University, Istanbul, Turkey, 2006

[21] Bozkurt, B., Dutoit, T.: An Implementation and Evaluation of Two-Diphone-based Synthesizers for Turkish, in Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Blair Atholl, Scotland, 2001, pp. 247-250

[22] Şayli, Ö., Arslan, L. M., Özsoy, A. S.: Duration Properties of the Turkish Phonemes, in Proc. 11th International Conference on Turkish Linguistics (ICTL 2002) Northern Cyprus, 2002

[23] Bozkurt, B., Ozturk, O., Dutoit, T.: Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection, in Proc. Eurospeech 2003, Geneva, Switzerland, 2003, pp. 277-280

[24] Oflazer, K., Inkelas, S.: A Finite State Pronunciation Lexicon for Turkish, presented at EACL Workshop on Finite State Methods in NLP, Budapest, Hungary, 2003

[25] Öztürk, Ö., Çiloğlu, T.: Modeling Segmental Duration For Turkish Text-To-Speech, in Proc. IEEE 12th Conference on Signal Processing and Communications (SİU-2004) Kusadasi, Turkey, 2004, pp. 272-275 (in Turkish)

[26] Arısoy, E., Arslan, L. M., Demiralp, M. N., Ekenel, H. K., Kelepir, M., Meral, H. M., Özsoy, A. S., Şayli, Ö., Türk, O., Can-Yolcu, B.: Duration of Turkish Vowels Revisited, presented at 12th International Conference on Turkish Linguistics (ICTL 2004) Izmir, Turkey, 2004

[27] Sak, H., Güngör, T., Safkan, Y.: Generation of Synthetic Speech from Turkish Text, presented at 13th European Signal Processing Conference (EUSIPCO 2005) Antalya, Turkey, 2005

[28] Türk, O., Schröder, M., Bozkurt, B., Arslan, L. M.: Voice Quality Interpolation for Emotional Text-to-Speech Synthesis, presented at 9th European Conference on Speech Communication & Technoloy (Interspeech 2005) Lisbon, Portugal, 2005

[29] Aşlıyan, R., Günel, K.: A Syllable-based Speech Synthesis System for Turkish Texts, presented at Akademik Bilişim (AB'08) Canakkale, Turkey (in Turkish) 2008

[30] Görmez, Z., Orhan, Z.: TTTS: Turkish Text-To-Speech System, in Proc. 12th WSEAS International Conference on Computers, Heraklion/Crete Island, Greece, 2008, pp. 977-982

[31] Yılmaz, A. E.: A Proposal of a Lexicon Set and Software Framework for Turkish Text-to-Speech Synthesis Applications, in Proc. IEEE 17th Conference on Signal Processing and Communications (SİU-2009) Side/Antalya, Turkey, 2009, pp. 956-959 (in Turkish)

[32] Oflazer, K., Inkelas, S.: The Architecture and the Implementation of a Finite State Pronunciation Lexicon for Turkish, Computer Speech and Language, 20(1) 2006, pp. 80-106

[33] Öğüt, F., Kiliç, M. A., Engin, E. Z., Midilli, R.: Voice Onset Times for Turkish Stop Consonants, Speech Communication, 48, 2006, pp. 1094-1099

[34] Sak, H, Güngör, T., Safkan, Y.: A Corpus-based Concenative Speech Synthesis System for Turkish, Turkish Journal of Electrical Engineering and Computer Sciences, 4(2) 2006, pp. 209-223

[35] Orhan, Z., Görmez, Z.: The Framework of The Turkish Syllable-based Concatenative Text-to-Speech System with Exceptional Case Handling, WSEAS Transactions on Computers, 7(10) 2008, pp. 1525-1534

[36] Yılmaz, A. E.: A Lexicon Set and Software Framework for Turkish Text-to-Speech Synthesis Applications, Journal of the Faculty of Engineering and Architecture of Gazi University, 24(4) 2009, pp. 735-744 (in Turkish)

[37] Demircan, Ö.: Phonological Order in Turkey Turkish – Phonemes in Turkey Turkish. Türk Dil Kurumu, Ankara, Turkey (in Turkish) 1979

[38] Selen, N.: The Science of Articulation and Acoustics – Turkey Turkish. Türk Dil Kurumu, Ankara, Turkey (in Turkish) 1979

[39] Ergenç, İ.: Spoken Language and Dictionary of Turkish Articulation. Multilingual, İstanbul, Turkey (in Turkish) 2002

[40] User, H.Ş.: Turkish Alphabet Systems throughout the History. Akçağ, Ankara, Turkey (in Turkish) 2006

[41] Crisan, M.: Chaos and Natural Language Processing, Acta Polytechnica Hungarica, 4(3) 2007, pp. 61-74