Studia Linguistica Universitatis Iagellonicae Cracoviensis 129 supplementum (2012)
DOI 10.4467/20834624SL.12.023.0807

KAMIL STACHOWSKI
Jagiellonian University, Cracow
kamil.stachowski@gmail.com

A PHONOLOGICAL ENCODING OF TURKISH FOR NEURAL NETWORKS

Keywords: experimental phonology, neural networks, Turkish

Abstract

The paper proposes a multi-dimensional, phonologically-aware numeric encoding of Turkish for use with neural networks. The system is evaluated and compared to *PatPho* (Li/MacWhinney 2002) in a test in which the network computes the shape of the past tense suffix.

0. Introduction

A number of attempts have been made to convert phonemes to numbers in a way that adequately renders the distances between them. See Li and MacWhinney (2002: 408f) for an overview. In the same paper, a new encoding is proposed, but it is not tested. At least for Turkish, it would seem, its effectiveness is limited.

The present paper proposes a different system (section 1.) and evaluates it using the Turkish past tense suffix, achieving a considerably higher accuracy (section 2.). Moreover, the proposed encoding is easier to port to different languages beacuse the method for chosing the specific numeric values is less arbitrary and, even more importantly, it is overt. Finally, the preliminary results of an actual application of the encoding are presented (section 3.).

Turkish is particularly well suited for testing an encoding of this kind for two reasons. Firstly, its phonology is relatively simple and symmetric, and so it serves well as a model which can be viewed as a minimal example, appropriate for an initial presentation. Secondly, Turkish has vowel harmony and consonant assimilations on morpheme boundaries, both of which are regular and entirely dependent on phonetics alone.

All the examples are presented phonologically in the Finno-Ugric transcription; see Stachowski K. (2011) for details. Unless specified otherwise, Turkish verbs are traditionally given in the infinitive, i.e. with the *-m$^{a}_{e}$k* suffix attached and separated with a dot.

1. Encoding

The system is inspired by *PatPho*, which was proposed in Li and MacWhinney (2002), but it departs considerably from the original. It, too, is based on three features (three dimensions). The meaning of one of the dimensions has been retained (Li/MacWhinney's 'D2' = my X axis) but the values on it have been rearranged to better reflect the human anatomy. The other two axes have been assigned linguistically more relevant meanings.

In this article, only an encoding for modern literary Turkish will be discussed. Vowel length, however, will be omitted because it has no impact on the shape of any of the Turkish suffixes and, therefore, is very difficult to test.

The system starts with a two-dimensional cross-section of the mouth. The X axis represents the place of articulation, and the Y axis the height of the channel between the tongue and the palate.

To accommodate for such features as voicedness or nasality, a third dimension has been added which represents the number of organs taking part in the articulation. Thus, voiceless consonants are at Z=1, voiced consonants and unrounded vowels at Z=2 (+ vocal cords), and nasal consonants, /l/, and rounded vowels at Z=3 (+ vocal cords + nasal cavity / sides of the tongue / lips).

Here, the scale is phonologically discrete rather than phonetically continuous, which appears to be a much more practical solution for any objectives that are not strictly phonetic. Represented are not sounds or even allophones, but phonemes.

Each step in all the dimensions has the value of 1.

Certain simplifications and assumptions have been made.

Bilabial and labiodental consonants are grouped together as 'labials'; by the same token, velars and glottals are all categorized as 'guttural'. These measures have been introduced in order to exclude near-empty categories (there are only two labiodental and one glottal consonant in Turkish).

Nasal consonants are classified as stops with an additional place of articulation, rather than as a separate manner of articulation as in *PatPho*. The pairs *k* : *ḱ*, *g* : *ǵ* and *l* : *l′* are considered allophonic. Their phonemic status is perhaps debatable in Turkish as a whole, but in the limited fragment tested here (170 verb stems, see 2.), they can be safely treated as allophones. Note that this approach is actually more demanding for the encoding as it provides the network with fewer clues about the harmonies of the tested stems.

Using the above system, Turkish phonemes can be visualized three-dimensionally as in fig. 1. or, perhaps more practically, in a flattened two-dimensional table as in

tab. 1. For example, /p/ is represented as an ordered triple (1, 1, 1) = (labial, stop, voiceless), /b/ as (1, 1, 2), /f/ as (1, 3, 1) etc.

Some insight into the relations between the phonemes in this encoding can also be obtained from the dendrogram in fig. 2. Note, however, that a two-dimensional representation inevitably distorts the three-dimensional space to a certain degree.

Many network architectures, including the perceptron used here, require the input data to be of a fixed length. This creates slots which have to be assigned a value even though from the linguistic point of view they are empty (e.g when shorter words are mixed with longer ones in the analyzed corpus, and need to be artificially lengthened). In the test, they were completed with dashes, which were subsequently converted to the triple (-5, -5, -5). The results suggest that this system is 'understandable' for the network. (See 2.2.)

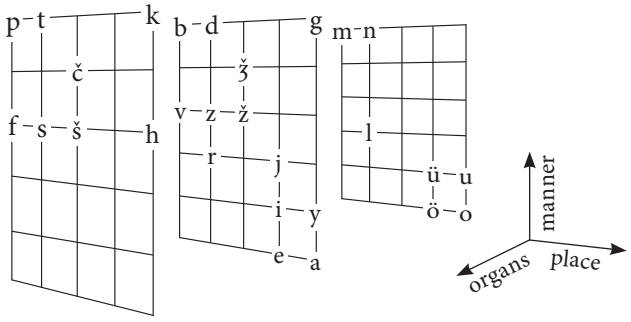

Figure 1. A three-dimensional representation of Turkish phonemes. See also tab. 1.

Feature	Labial	Alveolar	Post-alveolar	Palatal	Guttural	Value
Stop	p b m	t d n			k g –	**1**
Affricate			č ǯ –			**2**
Fricative	f v –	s z –	š ž –		h – –	**3**
Liquid		– r l		– j –		**4**
High vowel				– i ü	– y u	**5**
Low vowel				– e ö	– a o	**6**
	1	**2**	**3**	**4**	**5**	**Value**

Table 1. A flattened visualization of a three-dimensional representation of Turkish phonemes (fig. 1.). The first phoneme in each cell is at Z=1, the second at Z=2, and the third at Z=3. See also the resulting dendrogram in fig. 2.

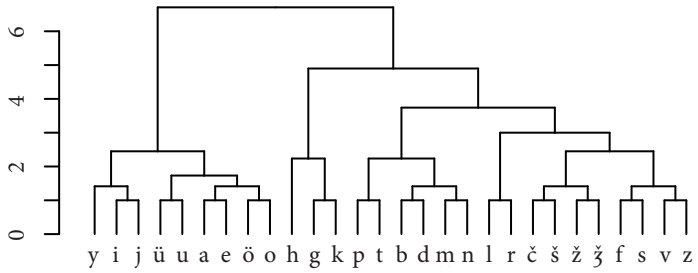

Figure 2. A dendrogram of the proposed encoding as in fig. 1. Note that a certain amount of bias is inevitable when a three-dimensional space is flattened to a two-dimensional picture.

The network outputs three numbers representing a point in a three-dimensional space, which is then interpreted as the nearest phoneme. The Canberra, Euclidean, Manhattan and maximum distances were tested; the average accuracies in the past tense suffix test (see 2.) were very similar for all four. The best results were identical: 1.0 for the training corpus and 0.976 for the test corpus. See tab. 2.

Distance	Training	Test	Total
Canberra	0.969013	0.771613	1.740627
Euclidean	0.969687	0.778893	1.748580
Manhattan	0.969687	0.778580	1.748267
Max. dist.	0.969687	0.768087	1.737773

Table 2. Average accuracies of the network in the past tense suffix test (see 2.) with different distance functions. The best results were identical for all four: 1.0 for the training corpus and 0.976 for the test corpus.

PatPho and the encoding proposed here differ in almost all the details but are nonetheless typologically similar in that they both exploit the existing linguistic knowledge in order to create a new computer model of a fragment of language.

Also the opposite approach has been attempted, perhaps even with more vigour. Computational methods have been employed to rediscover phonology in an unsupervised fashion, based solely on distribution analysis. To only name two, Rodd 1997 used recurrent neural networks and a phonologically blind encoding for Turkish, and obtained promising results, in particular in vowel harmony; Calderone 2009 used independent component analysis of phoneme collocationality to automatically identify phonological categories in English, Italian and Finnish. Both these studies, and others like them, are intriguing and appear to have the potential to serve as base for future encodings, with the added virtue of (at least greater) objectivity.

2. Test

To assay the effectiveness of the encoding, a test based on the Turkish past tense suffix was prepared. Its dictionary form is traditionally *-dy*, which is understood as any of the eight combinations of *d* ~ *t* + *y* ~ *i* ~ *u* ~ *ü*. *D* is chosen iff the final phoneme of the stem is voiced. *Y* and *u* are chosen iff the last vowel in the stem is back, and *u* and *ü* iff it is labial. Examples: *al.dy* '(s)he took', *et.ti* '(s)he did', *vur.du* '(s)he hit', *gör.dü* '(s)he saw'.

A hundred and seventy Turkish monosyllabic verb roots were collected, sorted alphabetically and the even roots used as the training corpus while the odd ones as the test corpus, in three different versions; see 2.2.

The test was run on a multi-layer perceptron as implemented in the *neuralnet* package for *R* (Fritsch/Günther 2012, see also Günther/Fritsch 2010). It was trained with the resilient backpropagation algorithm without backtracking (*rprop-*). For the tests with the encoding based on *PatPho*, the error threshold was set at 0.005; for the encoding proposed in this article, it was set at 0.1. The random seed was always set at 1. For all the other settings, the default values were kept.

2.1. *PatPho*

Unfortunately, Li and MacWhinney (2002) only give an encoding for English and do not explain in detail how they reached the exact numeric values for the different phonological features. The general idea appears to be sufficiently clear, however, and I tried to follow it as closely as possible when porting their system to Turkish.

The modifications are few and minor. For consonants, the only change was the removal of the *dental* feature from D2, because it is not represented in Turkish.

As for vowels, Turkish phonology is based entirely on three binary oppositions: frontness, roundness and height; see e.g. Pilancı, Demir and Yılmaz (2011: 24f). To reflect this, the features *central* and *mid* were removed from D2 and D3, respectively, and the features *mid-high* and *mid-low* in D3 replaced with *labial high* and *labial low*.

The numeric values were all preserved. See tab. 3. and the resulting dendrogram in fig. 3. The empty slots (see 1.) were represented by the triple (-1, -1, -1).

D1		D2		D3	
vowel	0.100	front	0.100	high	0.100
				labial high	0.185
		back	0.250		
				labial low	0.355
		bilabial	0.450	low	0.444
		labio-dental	0.528		
		alveolar	0.684	nasal	0.644

D1		D2		D3	
voiced	0.750	post-alveolar	0.762	stop	0.733
		palatal	0.841	fricative	0.822
		velar	0.921	approximant	0.911
voiceless	1.000	glottal	1.000	lateral	1.000

Table 3. A port of Li and MacWhinney's encoding to Turkish. See also the resulting dendrogram in fig. 3.

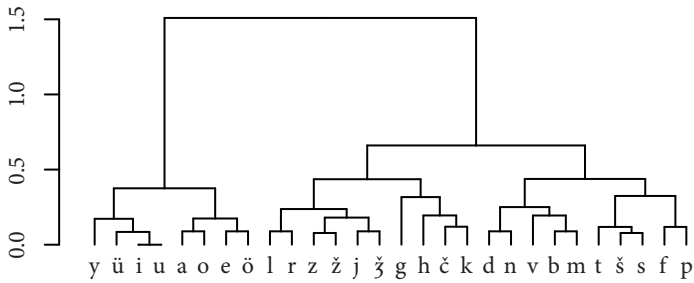

Figure 3. A dendrogram of an encoding for Turkish (tab. 3), based on Li and MacWhinney's (2002) proposition for English. Note that a certain amount of bias is inevitable when a three-dimensional space is flattened to a two-dimensional picture.

2.2. Comparison

As mentioned above, two encodings were tested on three versions of two corpora of 85 unique Turkish monosyllabic verb stems each. All words were two to four phonemes long.

In the first version, no specific template was used. Words of fewer than four phonemes were completed with dashes, represented by (-1, -1, -1) in the encoding based on *PatPho* and by (-5, -5, -5) in the one proposed in this article.

The second version ('CV-1') used a CVCC template filled from the left, i.e. with stems which only have one consonant in the auslaut never occupying the rightmost slot (e.g. `-al-` 'to take' or `bak-` 'to look'). The voicedness of the anlaut of the suffix is determined by the last phoneme of the stem. Voiceless consonants can co-occur with sonorants in Turkish, e.g. in *art.mak* 'to increase'. As a result, the choice between *d-* and *t-* in the suffix must be based on slot 4 if it is filled (e.g. `-art`), and on slot 3 if it is not (e.g. `bak-`).

The third version ('CV-2') used the same CVCC template, but filled from the right, i.e. with stems with a consonant in the auslaut always occupying the rightmost slot (e.g. `-a-l` 'to take' or `ba-k` 'to look'). With this method, the voicedness of the anlaut of the suffix can always be deduced from slot 4.

The test was run on networks with one to hundred hidden neurons. (See 2. above for detailed settings.) The results are presented in fig. 4. and tab. 4. Note that both encodings are typologically similar and their results can be directly compared without additional reservations.

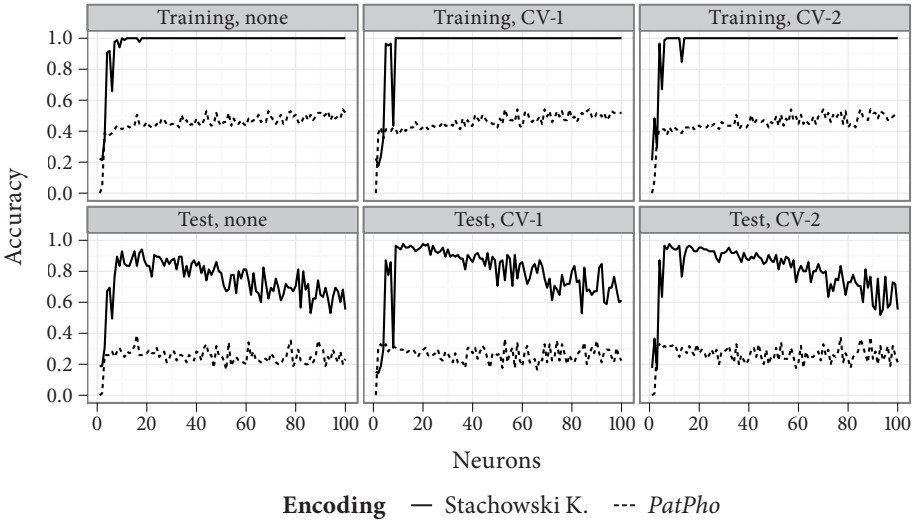

Figure 4. A comparison of the accuracy of neural networks in the Turkish past tense suffix test with an encoding based on *PatPho* (see 2.1.) and that presented in this article (see 1.). The plots are divided by corpora (rows) and templates used (columns). See tab. 4. for the exact numbers.

	Corpus/Result	*PatPho*	Stachowski K.
Average	Training	0.456	0.970
	Test	0.258	0.779
	Test, none	0.249	0.741
	Test, CV-1	0.263	0.791
	Test, CV-2	0.263	0.805
	Overall	0.357	0.874
Best	Training	0.541	1.000
	Test, none	0.388	0.941
	Test, CV-1	0.365	0.976
	Test, CV-2	0.376	0.976

Table 4. Accuracy of neural networks in the Turkish past tense suffix test with an encoding based on *PatPho* (see 2.1.) and that presented in this article (see 1.). See fig. 4. for an overview.

In total, the networks answered 51,000 questions for each encoding (85 words × 2 corpora × 3 templates × 100 different numbers of hidden neurons).

In 52% of the cases (54% in the test corpus), the Turkish port of *PatPho* returned a triple which in a three-dimensional space lay exactly between two phonemes. These answers were considered to be incorrect because it is hardly fair to expect the examiner to choose a single correct answer out of the many alternatives provided by the examinee. The best 'cheated' result on the test corpus would be 0.706 (53 neurons, CV-1).

With the encoding proposed in this article, this situation did not occur.

The best total result of 1.976 was achieved with four different settings: template CV-1 + 12/20/22 neurons, and template CV-2 + 8 neurons. In every case, the accuracy in the training corpus was 1.0, and 0.976 in the test corpus. The errors are shown in tab. 5.

Verb	CV-1, 12 neurons	CV-1, 20 neurons	CV-1, 22 neurons	CV-2, 8 neurons
čent.ti	*dl*	*ty*	+	*du*
jont.tu	*dü*	*ty*	+	*du*
art.ty	+	+	*žo*	+
sars.ty	+	+	*du*	+

Table 5. Errors of the four most accurate networks. Verbs are given with the past tense suffix (i.e. the correct answer) instead of in the infinitive.

The reasons for these errors are not obvious. *Čent.mek* and *jont.mak* are the only roots ending in *-nt* in either corpus, and *sars.mak* is the only one in *-rs*. As for *art.mak*, there are two more roots ending in *-rt* in the test corpus (*jyrt.mak* and *sürt.mek*) and three in the training corpus (*dürt.mek*, *ört.mek* and *tart.mak*). Generally, roots ending in a sonorant + voiceless consonant are relatively frequent in both corpora (10 in the training corpus and 9 in the test corpus), while the cluster sonorant + voiced consonant is not represented at all in the auslaut.

It should be also noted that six out of the eight errors are of a kind that beginners in Turkish very often make, too. Only *dl* and *žo* are completely incoherent.

3. Postscript

*-C-*type anlaut reduplication is a no longer productive method of intensification in the Turkic languages, whereby the initial (consonant and) vowel of the word is reduplicated and prepended to the base with a fixed 'closing consonant' in between, e.g. *ješil* 'green' → *je.m.ješil* 'completely green', *kara* 'black' → *ka.p.kara* 'pitch-black'.

In Turkish, the closing consonant can be one of *m*, *p*, *r* and *s*. The rules governing the choice are not known. Surely, they are not purely synchronic and phonetic, but this is nevertheless the approach that has dominated the research so far.

In particular, H.-G. Müller (2004) proposed a set of five (synchronic and phonetic) rules and in order to test them, asked 125 Turkish students to reduplicate three corpora: A with 100 words which do actually have reduplications in Turkish, B with 94 words which do not and C with 24 nonsense words (p. 251f). Next, he compared the results with his predictions – which, notabene, go against his own rules in about a third of the examples; see Stachowski K. [forthcoming].

A neural network was trained on the corpus Müller used to formulate his rules (a superset of his corpus A), and tested on his corpora B and C. With the random seed set at 1, the best coincidence with his predictions was achieved by networks containing 15, 22 and 35 hidden neurons. For each of these numbers, 125 networks were created, and on each occasion the seed was chosen randomly from a range of one to one million. The preliminary average results are given in tab. 6.

Corpus	Müller 2004	ANN, 15 neurons	ANN, 22 neurons	ANN, 6 neurons
A/Training	84.2%	76.43%	81.80%	65.86%
B	57.7%	54.21%	52.74%	57.56%
C	36.1%	32.46%	32.30%	35.00%

Table 6. A preliminary comparison of the average performance of different neural networks with the encoding proposed in this article, and the students interviewed by Müller (2004).

Taking into consideration that randomness is responsible for a significant part of the results in tab. 6., they still appear to be conspicuously similar. In itself, this is, of course, not a proof of the general feasibility of the encoding proposed in this article. I believe that it nevertheless does suggest the potential of this line of research.

References

Calderone B. 2009. Learning phonological categories by independent component analysis. – *Journal of Quantitative Linguistics* 16/2: 132–156.

Fritsch S., Günther F. 2012. neuralnet 1.31. cran.r-project.org/web/packages/neuralnet/index.html.

Günther F., Fritsch S. 2010. *neuralnet*: Training of neural networks. – *The R Journal* 2/1: 30–38.

Li P., MacWhinney B. 2002. PatPho: A phonological pattern generator for neural networks. – *Behavior Research Methods, Instruments & Computers* 34/3: 408–415.

Müller H.-G. 2004. *Reduplikationen im Türkischen. Morphophonologische Untersuchungen* [= *Turcologica* 26], Wiesbaden.

Pilancı H., Demir N., Yılmaz E. 2011. Türkçe Ses Bilgisi [= *T.C. Anadolu Üniversitesi Yayını* 2362 = *Açıköğretim Fakültesi Yayını* 1359], Eskişehir.

R Development Core Team 2011. *R: A language and environment for statistical computing*. Vienna.

Rodd J. 1997. Recurrent Neural-Network Learning of Phonological Regularities in Turkish. – Ellison T.M. (ed.) *CoNLL97: Computational Natural Language Learning*. Madrid: 97–106.

Stachowski K. 2011. Remarks on the usefulness of different types of transcription, with a particular regard to Turkic comparative studies. – *Suomalais-Ugrilaisen Seuran Aikakauskirja / Journal de la Société Finno-Ougrienne* 93: 303–338.

Stachowski K. [forthcoming] *Turkic -C-type anlaut reduplication*.