University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

CSE Journal Articles | Computer Science and Engineering, Department of

2019

Using Data Mining Algorithms to Discover Regular Sound Changes among Languages

Peter Revesz
University of Nebraska-Lincoln, prevesz1@unl.edu

Follow this and additional works at: https://digitalcommons.unl.edu/csearticles

Revesz, Peter, "Using Data Mining Algorithms to Discover Regular Sound Changes among Languages" (2019). *CSE Journal Articles*. 211.
https://digitalcommons.unl.edu/csearticles/211

This Article is brought to you for free and open access by the Computer Science and Engineering, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in CSE Journal Articles by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Using Data Mining Algorithms to Discover Regular Sound Changes among Languages

Peter Z. Revesz[1,a]

[1] *Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE, 68588, USA*

Abstract. This paper presents a method of using association rule data mining algorithms to discover regular sound changes among languages. The method presented has a great potential to facilitate linguistic studies aimed at identifying distantly related cognate languages. As an experimental example, this paper presents the application of the data mining method to the discovery of regular sound changes between the Hungarian and the Sumerian languages, which separated at least five thousand years ago when the Proto-Sumerian reached Mesopotamia. The data mining method discovered an important regular sound change between Hungarian word initial /f/ and Sumerian word initial /b/ phonemes.

1 Introduction

Regular sound changes between two languages indicate that they are cognate, that is, derive from a common ancestor [11]. A set of languages with a common ancestry is called a language family. Linguists studying various language families, for example the Indo-European language family, already found many examples of regular sound changes without the use computers [11]. Table 1 shows some examples from Pellard et al. [14]. In particular, the words in Table 1 illustrate the regular sound change from /b/ to /g/ between Greek and Sanskrit and the regular sound change from /g/ to /k/ between Sanskrit and Tokharian B.

Table 1. Examples of regular sound changes in Indo-European languages by Pellard et al. [14].

English	Greek	Sanskrit	Tokharian B
cow	boûs (βους)	gáv-	ke_u
come	baínō (βαινω)	gam-	käm-
-	barús (βαρυς)	gurú-	krā-mär

The abundance of cognate words, as exists for example between English and German, suggests a relatively recent separation of the two languages. In such cases, it is feasible to manually search for word pairs with regular sound changes. However, if two languages are only distantly related, then the search for regular sound changes becomes as difficult as looking for a needle in a haystack. Hence for more distantly related languages, the use of automated data mining techniques would become necessary to use.

In this paper we describe a data mining method for looking for cognate pairs of words in pairs of languages. We also apply the data mining method to the study of distant relationships between Hungarian and Sumerian languages. The Sumerian language is generally considered a language isolate, while Hungarian is classified to be a member of the Uralic language family. Nevertheless, Sumerian and Hungarian were already claimed to be cognate by Badiny [1], Baráth [2], Bobula [3], Csőke [4], Gosztony [6], Götz [7], Parpola [13], Tóth [24] and Zakar [27]. Sumerian and Tamil were also claimed to be cognate by Muttarayan [12]. However, Honti [9] pointed out that previous researchers did not find a satisfying set of regular sound changes between Hungarian and Sumerian. That situation contrasts greatly to the situation within the Uralic language family where many regular sound changes were already found [9].

Sumerian was spoken and written using a cuneiform script in Mesopotamia from around 3200 to 2000 BC. Although many historians trace the origin of Hungarians to the area north of the Caucasus Mountains, Hungarian is currently spoken mostly in present day Hungary and some neighboring countries in central Europe.

Therefore, there is a great temporal and geographic separation of Sumerian and Hungarian, which means that any possible relationship can only be a distant relationship. Therefore, the use of data mining algorithms could be very beneficial in this case.

The rest of this paper is organized as follows. Section 2 describes the data sources, including all the dictionaries used, and the representation of the input data. Section 3 presents the results of our data mining and a discussion of the results. Section 4 provides related works. Finally, Section 5 gives some conclusions and directions for future work.

2 Data sources

In this section, we first describe the dictionaries used in this study (Section 2.1), and then the representation of the input data (Section 2.2).

[a] Corresponding author: revesz@cse.unl.edu

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

Table 2. Examples of Sumerian and Uralic cognate words. The entry numbers in Parpola's dictionary are in the last column.

English	Hungarian	Other Uralic or Finno-Ugric	Sumerian	#
father	atya	ättä[Finnish]	ad-da	39
mother	anya	an[Zyrian] (husband's mother)	ama	102
hide (n.)	bőr (skin)	parva[Finnish] (leather coat), pěr[Khanty]	bar	259
drop, drip	csorog	ćork[Mansi], šoro[Finnish] (gurgle)	sur	2277
water	eső (rain) < esik (fall)	is[Mansi] (come down), äs[Selkup] (fall)	eš	715
tree	fa	puu[Finnish], pō[Selkup]	mu	1927
back, rear, tail	far	perä[Finnish]	bar	241
trim with axe	farag	pār[Mansi], pārge[Selkup]	bar	255
eye, face	fej (head)	uopi[Khanty] (look), vop[Mansi]	i-bí	1209
fear (v.)	fél	pěl[Khanty], pelkää[Finnish]	bu-luh	356
half, half-liquid	fél	päl[Mansi], pal[Udmurt]	bar	269
box, chest	fészek (nest)	pesä[Finnish]	pisaĝ	1998
blow (wind)	fúj	pŏγ[Khanty], pow[Mansi]	bu_7	346
drill	fúr	pura[Finnish]	bùru	379
braid, weave[1]	fon	păn[Khanty] (yarn), panne [Saami] (spin)	pan	1952
bend	fordul	porjal[Votyak] (spin)	bùru	377
wave	hab (foam)	kump[Khanty], kop[Mansi]	gúb (snow)	867
destroy	hal (die)	kăla[Khanty], kāl[Mansi], koule[Finnish] (die)	hulu	1164
fish	hal	koule[Finnish], kole[Nganasan], kul[Zyrian]	ku_6	1423
walk, go	halad	koγel[Khanty], kulke[Finnish]	kul	1446
boy	here (scrotum)	kar[Khanty] (male)	ĝuruš	1092
raven, eagle[1]	holló	kolāk[Mansi] kulé[Selkup]	hurin[1]	1192
length measure	hosszú (long)	košew[Mansi] (long), kuź[Zyrian] (length)	_éše	712
lie down	huny (rest, close eye)	kŏń[Khanty], koń[Mansi] (close eyes)	huna	1183
two	két	kit[Mansi], kaksi[Finnish]	kad	1300
stone	kő	kaw[Mansi]	kín	1392
sinew	_ín	tɛn[Mansi], suoni[Finnish]	sa	2054
piece	mar (bite)	murta[Finnish] (break)	mir	1083
what	mit ('t' is accus. suffix)	mitä[Finnish], mida[Estonian]	_ta	2460
wash (hand)	mos (wash) > mosdik	moška[Mari]	maš (purify)	1654
woman, bride[1]	nő, néné (elder sister)	nī[Mansi], naine[Estonian]	nu-nus	1917
heart	szív	sěm[Khanty], šäm[Mansi]	ša-ab	2286
to fly	toll (feather)	tēl[Mansi], to[Yurak] (feather, wing)	dal	425
be wide	vas (iron) > vastag (thick)	vaski[Finnish] (copper), baza[Kamas] (iron)	peš	1961

2.1 Sumerian and Hungarian dictionaries

Parpola's *Etymological Dictionary of the Sumerian Language* [13] describes the Uralic etymologies for over three thousand Sumerian words. Table 2 from Revesz [21] shows some of the etymologies from Parpola's dictionary.

In Table 2 the first column is the equivalent English word or a short description of the meaning in English, the second, third and fourth columns give the Hungarian, Uralic and Sumerian cognate words, respectively. The last column is the entry number of the Sumerian word from Parpola's dictionary. The third column is a combination of Parpola and Zaicz [27]. The language in which a word occurs is indicated as a superscript but is omitted when it is obvious which language we discuss.

In Table 2 we highlighted in red the corresponding consonants. For example, in the first row the Hungarian consonant /ty/ corresponds to the geminate consonants /tt/ in Finnish and /dd/ in Sumerian. In addition to Parpola's dictionary, we also considered the ePDS, the online version of the *Pennsylvania Dictionary of Sumerian* [22]. We found many possible cognate Sumerian and Hungarian word pairs using the ePDS and Zaicz [27].

2.2 Representation of the input data

After the data collection described in Section 2.1, we represented all pairs of Hungarian and Sumerian cognate words by an ARFF file as shown in Fig. 1. The ARFF file uses six attributes. The first three attributes are for the initial, medial and final consonants of the Hungarian word, while the next three attributes are for the initial, medial and final consonants for the Sumerian word. The attribute values are the set of consonants that occur in Hungarian or in Sumerian or the special value "empty" that denotes the omission of a consonant. Since all the words had at most three consonants, the above description gave a complete representation of the consonant base of each word. For the Sumerian words we used the phonetic reconstructions of Parpola [13], and for the Hungarian words we relied on the phonetics given in Zaicz [27].

```
% Title: Database for identifying regular sound changes between Hungarian and Sumerian
@relation soundchange
%First three attributes are for Hungarian, and the second three attributes are for Sumerian words.
@attribute init-cons1      {empty, b, c, ch, d, f, g, h, j, k, l ,m ,n ,p, q, r, sh, s, t, ty,v, w, z}
@attribute medial-cons1 {empty, b, c, ch, d, f, g, h, j, k, l ,m ,n ,p, q, r, sh, s, t, ty,v, w, z}
@attribute final-cons1     {empty, b, c, ch, d, f, g, h, j, k, l ,m ,n ,p, q, r, sh, s, t, ty,v, w, z}
@attribute init-cons2      {empty, b, c, ch, d, f, g, h, j, k, l ,m ,n ,p, q, r, sh, s, t, v, w, z}
@attribute medial-cons2 {empty, b, c, ch, d, f, g, h, j, k, l ,m ,n ,p, q, r, sh, s, t, v, w, z}
@attribute final-cons2     {empty, b, c, ch, d, f, g, h, j, k, l ,m ,n ,p, q, r, sh, s, t, v, w, z}
@data
empty,ty,empty,  empty,d,empty
empty,n,empty,  empty,m,empty
b,empty,r,      b,empty,r
ch,r,g,         s,r,empty
ch,empty,r      s,empty,r
empty,s,empty,  empty,sh,empty
f,empty,empty,  m,empty,empty
f,empty,r,      b,empty,r
f,r,g,          b,r,empty
f,j,empty,      b,empty,empty
f,empty,l,      b,l,empty
f,empty,l,      b,r,empty
f,empty,t,      p,s,empty
f,empty,j,      b,empty,empty
f,empty,r,      b,r,empty
f,empty,empty,  p,empty,n
f,empty,empty,  b,r,empty
h,empty,b,      g,empty,b
h,empty,empty,  h,empty,l
h,empty,empty,  k,empty,empty
h,l,empty,      k,l,empty
h,r,m,          empty,m,sh
h,r,empty,      g,r,empty
h,empty,l,      h,r,empty
```

Figure 1. The first few lines of the ARFF file representing the initial, medial and final consonants within the pairs of cognate Hungarian and Sumerian words.

For example, for the cognate word pair of Hungarian *atya* and Sumerian *adda*, which we saw in the first row of Table 2, the initial and the final consonants are missing while the medial consonants are /ty/ and /dd/, respectively. Hence the first line of data, that is, "(*empty, ty, empty, empty, d, empty*)" as shown in Fig. 1 represents the consonants in this pair of words. There were a total of 177 records in our data.

```
 1. medial-consl=r final-cons2=empty 16 ==> medial-cons2=r 16    <conf:(1)> lift:(5.36) lev:(0.07) [13] conv:(13.02)
 2. init-cons2=b final-cons2=empty 9 ==> init-consl=f 9    <conf:(1)> lift:(10.41) lev:(0.05) [8] conv:(8.14)
 3. medial-cons2=g 10 ==> medial-consl=g 9    <conf:(0.9)> lift:(11.38) lev:(0.05) [8] conv:(4.6)
 4. medial-consl=r 25 ==> medial-cons2=r 22    <conf:(0.88)> lift:(4.72) lev:(0.1) [17] conv:(5.08)
 5. init-consl=empty final-consl=empty 12 ==> final-cons2=empty 10    <conf:(0.83)> lift:(1.55) lev:(0.02) [3] conv:
 6. init-cons2=m 16 ==> init-consl=m 13    <conf:(0.81)> lift:(8.46) lev:(0.06) [11] conv:(3.62)
 7. init-cons2=b 15 ==> init-consl=f 12    <conf:(0.8)> lift:(8.33) lev:(0.06) [10] conv:(3.39)
 8. init-consl=m 17 ==> init-cons2=m 13    <conf:(0.76)> lift:(8.46) lev:(0.06) [11] conv:(3.09)
 9. init-consl=f final-cons2=empty 12 ==> init-cons2=b 9    <conf:(0.75)> lift:(8.85) lev:(0.05) [7] conv:(2.75)
10. init-consl=f init-cons2=b 12 ==> final-cons2=empty 9    <conf:(0.75)> lift:(1.4) lev:(0.01) [2] conv:(1.39)
```

Figure 2. The ten rules found by the Weka association rule data miner.

3 Experimental results

3.1 Association rules found

We used an association rule data mining [15] algorithm that was implemented within the Weka system. The association rule data mining learns association rules given as input data a number of *itemsets*. In typical applications, the itemsets are the set of items that are purchased together by customers. If they are purchased together by a large number of customers, then their association has a large support. The main motivation for association rule data mining was that if a customer purchased some items, then other items that are frequently associated with the purchased items could be suggested to the customer.

Our application of association data mining moves well beyond the original intended application, but it is still very intuitive. If a strong association is found between two different Hungarian and Sumerian consonants in the same (initial, medial or final) position, then it indicates a regular sound change between those two consonants.

We had to experiment with different parameters for the association rule data mining. We used minimum metric (or confidence) = 0.7 and minimum support = 0.05, which required nine instances supporting the rule found. With these parameters, the Weka association rule data miner found the ten best rules shown in Fig. 2. The non-trivial rules, where there was an actual sound change, were rules 2, 7, 9, and 10. However, these four rules are just minor variations of the following main rule:

$$\text{init-cons1} = \text{f} \rightarrow \text{init-cons2} = \text{b} \qquad \text{(I)}$$

The above rule means that if the Hungarian initial consonant is /f/, then the Sumerian initial consonant is /b/. We can find the following examples of Rule (I) in the input database:

fal$^{\text{Hungarian}}$	~ ba-ar$^{\text{Sumerian}}$
far$^{\text{Hungarian}}$	~ bar$^{\text{Sumerian}}$
farag$^{\text{Hungarian}}$	~ bar$^{\text{Sumerian}}$
fél$^{\text{Hungarian}}$	~ bar$^{\text{Sumerian}}$
fél$^{\text{Hungarian}}$	~ bu-luh$^{\text{Sumerian}}$
fokos$^{\text{Hungarian}}$	~ bulug$^{\text{Sumerian}}$
fordul$^{\text{Hungarian}}$	~ bùru$^{\text{Sumerian}}$
fúl$^{\text{Hungarian}}$	~ bulug$^{\text{Sumerian}}$
fürt$^{\text{Hungarian}}$	~ buru$_{14}$$^{\text{Sumerian}}$

3.2 Discussion of the results

Although the association rule data mining found nine examples of /f/ and /b/ correspondences, it has overlooked the following /f/ and /p/ correspondences, which had only a five instance support:

fa$^{\text{Hungarian}}$	~ pa$^{\text{Sumerian}}$
fejsze$^{\text{Hungarian}}$	~ pa-a-šu$^{\text{Sumerian}}$
fészek$^{\text{Hungarian}}$	~ pisaĝ$^{\text{Sumerian}}$
fog$^{\text{Hungarian}}$	~ pag$^{\text{Sumerian}}$
fon$^{\text{Hungarian}}$	~ pan$^{\text{Sumerian}}$

Clearly, the above set of instances cannot be all ignored. What broader context is that the Hungarian word initial /f/ corresponds to the Sumerian word initial /b/ if the Sumerian medial consonant is a liquid /l or /r/ but it corresponds to the word initial /p/ otherwise. That can be summarized by the following association rules:

$$\text{init-cons1} = \text{f}, \text{medial-cons2} = \text{l} \rightarrow \text{init-cons2} = \text{b} \qquad \text{(II)}$$

$$\text{init-cons1} = \text{f}, \text{medial-cons2} = \text{r} \rightarrow \text{init-cons2} = \text{b} \qquad \text{(III)}$$

$$\text{init-cons1} = \text{f}, \ \text{medial-cons2} \neq \text{l}, \text{medial-cons2} \neq \text{r} \rightarrow \text{init-cons2} = \text{p} \qquad \text{(IV)}$$

4 Related works

Revesz [21], an earlier, manual attempt to find regular sound changes between Hungarian and Sumerian, already described sound change rules (II), (III) and (IV). The regular sound change rules show that Hungarian and Sumerian are cognate languages. In addition, Revesz [21] classified the *Euphratic* language, which is a proto-Sumerian language or substrate according to Whittaker [25], into the *West-Ugric* branch of the Uralic language family, which is within the Ugric branch together with the *Ob-Ugric* branch [9] that contains the Khanty and Mansi languages now spoken in Northwestern Siberia.

According to the recent translations of the Minoan Linear A script [20], the Cretan Hieroglyphic script [18, 19], and the Phaistos Disk [17], the Minoan language can be also classified as West-Ugric. Moreover, the Minoan scripts show some similarities to the Old Hungarian script [16], which is also called *rovásírás* in Hungarian and also written sometimes as Rovas in English language publications.

However, similarity of scripts is not a proof of similarity of language because some scripts could be widely adopted and used to write languages that belong to different language families. Indeed, members of the Cretan Script Family [16], which includes Cretan Hieroglyphs, Linear A, Old Hungarian, as well as Linear B, the Carian alphabet, and Tifinagh, all adopted the same ancient script.

5 Conclusions and future work

We presented a method of using association rule data mining algorithms for the automatic discovery of regular sound changes between a pair of languages. In the future we plan to expand this work to consider more than two languages. For example, in Table 1 compares four languages. Since there are only three rows none of the discovered associations would have more than three itemsets as support. That may be considered too low between a pair of languages. However, when four languages exhibit a complex regular sound change as shown in Table 1, then the association found can be considered to be a strong evidence. The reason is that each row of Table 1 is equivalent to six different pairs of languages. Hence it makes sense to apply data mining simultaneously to all languages within a language family.

Regular sound changes also need to be studied to together with grammar with the aim of discovering novel similarities between the Hungarian [11] and the Sumerian grammars [5]. Finally, Rules (I-IV) can be expressed as constraint database rules [10], whose implications can be also studied using computer algorithms.

References

1. F. J. Badiny, New lines for a correct Sumerian phonetics to conform with the cuneiform scripts, *Proc. 29th Int. Congress of Orientalists*, (1973)
2. T. Baráth, *A Magyar Népek Őstörténete (The Prehistory of the Hungarian People)*, Vols. 1-5, (Somogyi Publ., Franklin Park, New York, USA, 1974)
3. I. Bobula, *Sumerian affiliations; A plea for reconsideration*, manuscript, (Washington D.C., USA, 1951)
4. S. Csőke, *Szumir-Magyar Egyeztető Szótár (Sumerian-Hungarian Correspondences Dictionary)*, (Turáni Akadémia Publishing, Buenos Aires, Argentina, 1974)
5. D. A. Foxvog, Introduction to Sumerian Grammar, manuscript, (University of California at Berkeley, 2012)
6. K. Gosztony, *Dictionaire D'etymologie Sumerienne et Grammaire*, (De Boccard, Paris, France, 1975)
7. L. Götz, *Keleten Kél a Nap (The Sun Rises on the East)*. (Püski Publishing., Budapest, Hungary, 1989)
8. L. Honti, Characteristic features of Ugric languages (observations on the question of Ugric unity), *Acta Linguistica Academiae Scientiarum Hungaricae*, **29,** 1/2, pp. 1-26, (1979)
9. L. Honti, ed., *A Nyelvrokonságról - Az török, Sumer és Egyéb Áfium Ellen Való Orvosság,* (Tinta Publishing, Budapest, Hungary, 2010)
10. P.C. Kanellakis, G.M. Kuper, P.Z. Revesz, J. of Comp. and Sys. Sciences, **51**, 1, pp. 26-52, (1995).
11. J. Kiss and F. Pusztai, eds., *A Magyar Nyelvtörténet Kézikönyve* (Tinta Publ., Budapest, Hungary, 2018)
12. K. L. Muttarayan, Sumerian, Tamil of the First Cankam, *J. of Tamil Studies,* **8**, pp. 40-61, (1975)
13. S. Parpola, *Etymological Dictionary of the Sumerian Language, Vols. 1 and 2,* (Foundations for Finnish Assyriological Research, Helsinki, Finland, 2016)
14. T. Pellard, L. Saquant, and G. Jacques, L'indo-européen n'est pas un mythe, *Bulletin de la Société de Linguistique de Paris*, Peeters Publishers, **113** (1), pp. 79−102, (2018)
15. P. Z. Revesz, *Introduction to Databases: From Biological to Spatio-Temporal*, (Springer, New York, 2010)
16. P.Z. Revesz, Bioinformatics evolutionary tree algorithms reveal the history of the Cretan script family, Int. Journal of Applied Mathematics and Informatics, **10**, pp. 67-76, (2016)
17. P.Z. Revesz, A computer-aided translation of the Phaistos Disk, Int. Journal of Computers, **10**, pp. 94-100, (2016)
18. P.Z. Revesz, A computer-aided translation of the Cretan Hieroglyph script, Int. Journal of Signal Processing, **1**, pp. 127-133 (2016)
19. P.Z. Revesz, A translation of the Arkalochori Axe and the Malia Altar Stone, WSEAS Transactions on Information Science and Applications, **14,** 1, pp. 124–133, (2017)
20. P.Z. Revesz, Establishing the West-Ugric language family with Minoan, Hattic and Hungarian by a decipherment of Linear A, WSEAS Transactions on Information Science and Applications, **14,** 1, pp. 306–335, (2017)
21. P.Z. Revesz, Sumerian contains Dravidian and Uralic substrates associated with the Emegir and Emesal dialects, *WSEAS Transactions on Information Science and Applications*, **16,** 1, pp. 8-30, (2019)
22. *The Pennsylvania Sumerian Dictionary*, Available: http://psd.museum.upenn.edu
23. M.-L. Thomsen, *The Sumerian Language: An Introduction to its History and Grammatical Structure*, (Akademisk Forlag, Copenhagen, Denmark, 1984)
24. A. Tóth, *Hungarian, Sumerian and Egyptian. Hungarian, Sumerian and Hebrew: Two Addenda to the Etymological Dictionary of Hungarian,* (Mikes International, Hague, Netherlands, 2007)
25. G. Whittaker, The case for Euphratic, *Bulletin of the Georgian National Academy of Sciences*, **2**, 3, pp. 156-168, (2008)
26. G. Zaicz, chief editor, *Etimológiai Szótár: Magyar Szavak és Toldalékok Eredete,* (Tinta Publishing, Budapest, Hungary 2006).
27. A. Zakar, Sumerian-Ural-Altaic Affinities, *Current Anthropology*, **12**, 2, pp. 215-216, (1971)