Constructing a Philippine Language Family Tree by Clustering Character Trigrams

Angelica H. Dela Cruz, Maria Cristina P. Co, Adrian Martin S. Sy, and Nathaniel Oco
National University
delacruzah@yahoo.com, mariacristinaco80@gmail.com, adrian1009sy@gmail.com, nathanoco@yahoo.com

ABSTRACT

Proper discernment between closely related languages is one problem in language identification. Strong similarities between certain languages could result in lower recall rates. In this paper, we clustered 43 Philippine languages by using trigrams to identify closely-related languages, and constructed a language family tree that conveys the origin of the languages. We collected online religious text documents and used 100,000 words for the 42 domain languages as training data. For the Yami language, a dictionary project, short story and a few transcribed verses were used. These were cleaned and character trigrams were generated. The languages were clustered using two algorithms, farthest first algorithm and simple k-means algorithm.

Dice's Coefficient on trigram profiles was used as metric for language similarity to validate the results. For farthest first algorithm, 56% of the domain languages were clustered the same based on its values, and for simple k-means algorithm, 74% were clustered the same. The language family tree constructed were compared to a Philippine language relations map and Philippine Language Family Tree of Ethnologue. The results of the experiment showed that language similarity plays an important role in language clustering. For future work, geographic location and phonetic alphabet of the languages will be explored and used as features.

Categories and Subject Descriptors

G.5.5 [**Information Systems**]: Information Retrieval – *Clustering and Classification*

General Terms

Languages

Keywords

Language clustering, language similarity, language family tree, trigrams, Dice's coefficient, Philippine languages

1. INTRODUCTION

Language clustering is grouping the languages based on their similarities and relatedness to each other. This approach is used in phylogenetic analysis. Phylogenetic methods are used in constructing language family trees that convey the origin of a certain language. Features may include lexical, phonological and morphological information [1].

Existing studies used trigram models in clustering languages [2], [3] some used trigram rankings to identify language similarities [4]. Trigrams are models that are commonly used in language modeling as the value of 3 is enough to represent the unique character sequences of a language while covering single-letter words [5]. For example, the word "cluster" will produce a trigram model of {"_cl", "clu", "lus", "ste", "ust", "ter" , "er_"}. The purpose of this study is to construct a language family tree of Philippine languages by using trigrams to cluster 43 languages (see Table 1), a feat which has never been done automatically before.

Yami language, a language of Taiwan was also included in the domain languages because it was reportedly similar to the Ivatan language of the Philippines according to Ethnologue,

This paper is organized as follows: related studies in section 2, methodology in section 3, results and evaluation in section 4, and conclusion and further work in section 5.

Table 1. Philippine Languages and language code

No.	Language	Language Code[1]	No.	Language	Language Code
1	Agta	agt	23	Kagayanen	cgc
2	Agutaynen	agn	24	Kalinga	kyb
3	Alangan	alj	25	Kallahan	kak
4	Ayta	sgb	26	Kapampangan	pam
5	Balangao	blw	27	Kinaray-a	krj
6	Bikol	bik	28	Maguindanao	mdh
7	Binukid	bkd	29	Manobo	mta
8	Blaan	bpr	30	Mansaka	msk
9	Bolinao	smk	31	Maranao	mrw
10	Bontok	lbk	32	Masbatenyo	msb
11	Buhid	btw	33	Matigsalug	mbt
12	Cebuano	ceb	34	Pangasinense	pag
13	Chavacano	cbk	35	Paranan	prf
14	Hanunoo	hnn	36	Sama	sml
15	Hiligaynon	hil	37	Sambal	xsb
16	Ifugao	ifk	38	Tagalog	tgl
17	Ilocano	ilo	39	Tausug	tsg
18	Inabaknon	abx	40	Tiruray	tiy
19	Iranun	ilp	41	Waray	war
20	Iraya	iry	42	Yami	tao
21	Itawit	itv	43	Yakan	yka
22	Ivatan	ivv			

2. RELATED STUDIES

2.1 Language Clustering

Clustering is an unsupervised learning that finds natural grouping of instances given unlabeled data. It is a process of grouping data based on its similarities. Existing studies have clustered and measured similarity of languages using trigram models. These studies include the use of trigram models in classifying and clustering different Philippine languages by implementing a language identification system [2] and used the results to identify language clusters using farthest first algorithm. Another study [6] implemented a trigram-based language identification system for 20 Philippine languages, identified the languages and generated

[1] http://www.ethnologue.com/country/PH/languages

Philippine Computing Journal Dedicated Issue on Natural Language Processing, pages 10–17
Vol. XI No. 1 August 2016

clusters of languages using simple k-means clustering. Simple k-means clustering is also used in studies that used trigram ranking as metric for language similarity and clustering [4]. The presence and absence of trigrams were used by a study [3] to cluster 19 different languages using hierarchical and k-means algorithm.

2.2 Phylogenetic Trees

Phylogenetic methods are used to build evolutionary trees of languages. Linguistic phylogenetic trees convey the evolution of a language family. The family tree can be constructed on the basis of characteristics that are common to sets of languages. This includes lexical, phonological, and morphological affinities [1].

Traditional methods of constructing phylogenetic trees are typically based on lexical and phonological data ignoring information from any other level of analysis. Cluster-based methods produce similarity trees by computing the distance scores between languages directly from the corpus data and these are based from phonological and syntactic features [13]. Traditional methods do not consider a linkage criterion while cluster-based methods consider the relationship between languages based on the features.

Some studies that involve constructing phylogenetic trees and comparing it to other existing phylogenetic trees include [7] that used different corpus-based measures then compared the trees obtained.

Another study [8] used model-based method and vector-based method in reconstructing language family trees of Indo-European languages from non-native English texts. Agglomerative hierarchical clustering were used and it is stated in the study that statistical methods are used for measuring similarity of languages and a proposed method for constructing language family trees using clustering. Phylogenetic analyses are used by some researchers in learning the ancient history of languages. These studies include [9] the use of a Bayesian computational phylogenetic analysis of semitic languages (oldest written languages) that identifies an early bronze age origin of semitic in the near east and a study [10] that shows statistical phylogenetic analysis supports the traditional steppe hypothesis about the origins and dispersal of Indo-European language family and also confirms the reliability of statistical inference of reconstructed chronologies. It also includes a study [14] that reported the results of a quantitative analysis of lexical similarity between some languages of Tibeto-Burman and Austro-Asiatic to create phylogenetic trees. Reference [13] explores the relationship between the generic trees that are based from lexical and phonological features of the language and the similarity trees that are based on the phonological and syntactic features of the language that can be directly computed from corpus data.

For this paper, we worked on these concepts as inspiration. N-grams of size three or trigrams were used because higher values for *n* cannot cover single-letter words and lower values are not enough to represent unique character sequences of a language. Trigrams are enough to cover these [5]. In order to validate the clusters made, Dice's Coefficient [5] was used as metric for evaluating language similarity.

3. METHODOLOGY

3.1 Collection and Cleaning

Religious text documents were collected online and were used as training data for the 42 domain languages. These documents were used because the bible is already translated in different languages and can be accessed publicly. For the Yami language, an online dictionary project, a short story of the language and a few transcribed verses of The Yami New Testament ("Seysyo No Tao. Avayo A Seysyo") were used as the training data.

The number of words collected ranges from 163,000-306,000 depending on the language. For the training data to have an equal size, 100,000 words were used. This is important to have fair results with the languages considering the resources gathered. But because there are only limited resources for the Yami language, only 14,000 words were used as training data.

According to [12] a corpus must be representative in order to be appropriately used as the basis for generalizations concerning a language as a whole. The size of the corpora is not the only factor to consider in order for a corpus to be representative. A 1,000-word sample are already reliable for corpus with common features. The representativeness of a corpus depends on what texts are included and excluded from the corpora, the text categories (genre) included in the corpora and the distribution of features in the corpora.

In this study, the size of the gathered corpora was statistically representative because the gathered corpora for the languages were all in the same category (genre) and the size were already reliable for the common features.

The text documents were cleaned to remove all the unnecessary characters in generating trigrams. These were cleaned by utilizing regular expressions (Table 2) and removing special characters and English words included in the documents using Notepad++[2].

Table 2. Regular expressions used in cleaning

Find	Replace	Effect on Corpus
[“”?.,;:!()]	_(Space)	Replace all punctuation marks and quotation marks into a white space
[0-9]	_(Space)	Replace all numbers into a white space

3.2 Data Processing

After cleaning the data, the trigram profiles were generated. Apache Nutch[3] was used and each trigram profile contains the top 1,000 trigrams. The top 10 trigrams per language are shown in Table 3 (see appendix).

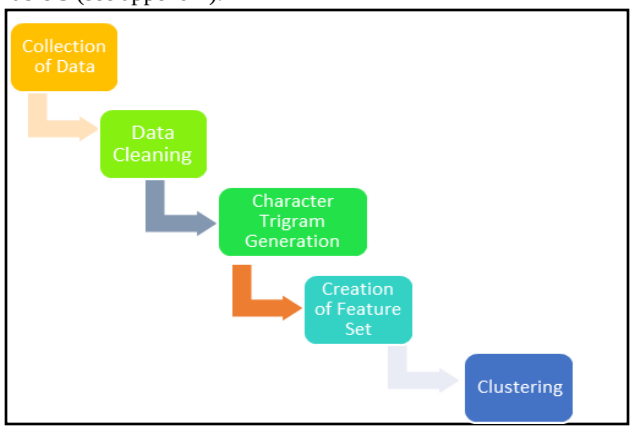

Figure 3-1. Clustering Procedure

[2] https://notepad-plus-plus.org/

[3] http://nutch.apache.org/

A total of 43,000 trigrams of all the domain languages were placed in a spreadsheet application and the data were filtered to produce a total of 4,598 unique trigrams. These unique trigrams were used to create representations of data per language. The data were fed to Weka, a data mining tool by the University of Waikato afterwards to automatically cluster the languages.

4. RESULTS AND EVALUATION

4.1 Clusters

After the data was fed to Weka, language clusters were identified using farthest first and simple k-means clustering algorithms. The number of clusters was set to 8 based on the grouping of the domain languages according to Ethnologue. Euclidean distance was set as the distance function for simple k-means.

Below are the results using farthest first algorithm.

- Cluster 0: Agta, Agutaynen, Ayta, Bikol, Binukid, Bolinao, Cebuano, Hanunoo, Hiligaynon, Ifugao, Ilocano, Inabaknon, Itawit, Kagayanen, Kalinga, Kallahan, Kapampangan, Kinaray-a, Maguindanao, Manobo, Mansaka, Masbatenyo, Matigsalug, Pangasinense, Paranan, Tagalog, Tausug, Tiruray, Waray, Yakan
- Cluster 1: Sama
- Cluster 2: Blaan
- Cluster 3: Chavacano
- Cluster 4: Yami and Ivatan
- Cluster 5: Iranun
- Cluster 6: Balangao and Bontok
- Cluster 7: Alangan, Buhid, Iraya, Maranao, Sambal

The results showed that Cluster 0 has the most number of closely related languages with 30 languages. Cluster 7 with 5 languages grouped as similar languages and Clusters 1, 2, 3, 4, 5 and 6 are clusters of languages that are considered outliers. Outliers are languages having very few members in a group or the only language in a group.

Below are the actual clusters based on the results of using simple k-means algorithm:

- Cluster 0: Ayta, Bolinao, Cebuano, Chavacano, Hiligaynon, Inabaknon, Itawit, Kapampangan, Kinaray-a, Masbatenyo,Paranan, Tagalog, Tiruray, Waray
- Cluster 1: Agutaynen, Alangan, Iraya, Ivatan, Maranao, Sambal, Yami
- Cluster 2: Bikol, Ilocano, Kagayanen, Kalinga, Mansaka, Pangasinense, Sama
- Cluster 3: Buhid
- Cluster 4: Balangao and Bontok
- Cluster 5: Ifugao and Kallahan
- Cluster 6: Binukid, Hanunoo, Maguindanao, Tausug
- Cluster 7: Agta, Blaan, Iranun, Manobo, Matigsalug, Yakan

The results showed that Clusters 3, 4 and 5 are the clusters considered having outlier languages. The rest of the languages are clustered together with the other languages closely-related to them.

The use of two different algorithms, farthest first and simple k-means showed 2 different results of clusters of the domain languages. As shown in the results, Yami and Ivatan language are similar languages as they are grouped in one cluster for both results.

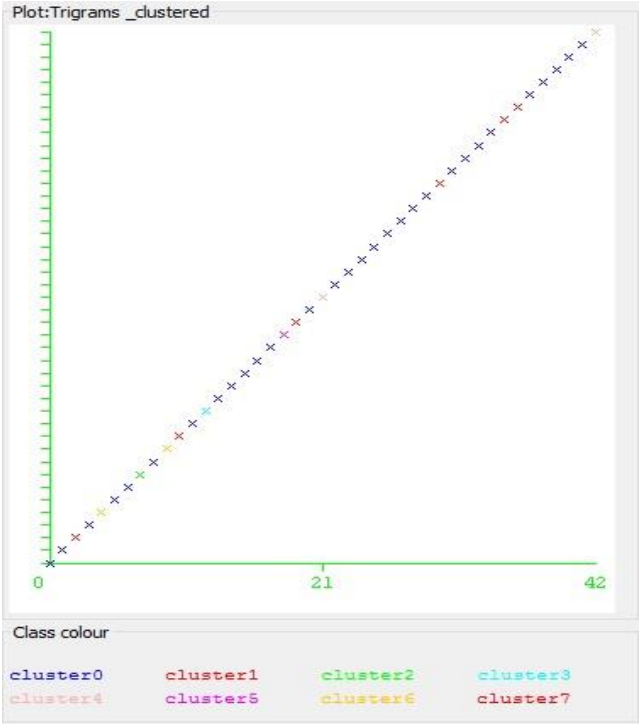

Figure 4-1. Visualized Cluster Assignments for Farthest First Algorithm

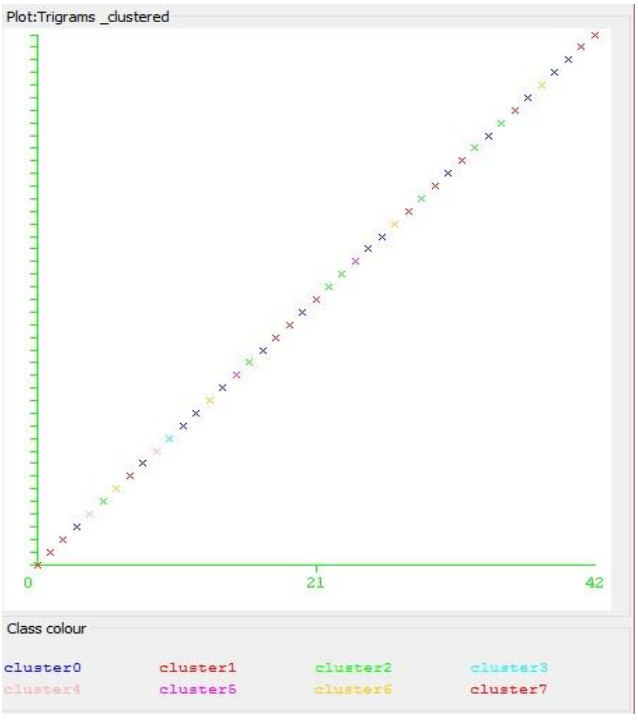

Figure 4-2. Visualized Cluster Assignments for Simple k-means Algorithm

The constructed language family tree for farthest first and simple k-means algorithm are shown in Figures 8-1 and 8-2 (see appendix)

4.2 Evaluation

In order to evaluate the results, DCTP (Dice's Coefficient on Trigram Profiles) was used as the metric for language similarity. The higher the value, the more similar the languages are.

$$Dice'sCoefficient = 2(X \cap Y)/(X + Y)$$

Equation 1. Dice's Coefficient Formula

Dice's coefficient as defined in Equation 1 was used to evaluate the clusters made, where X and Y are two different trigram profiles of two different languages. The results of the evaluation using Dice's Coefficient as metric is shown in Table 4 (at the last page). The DCTP values used in evaluating the results were limited to 2 decimal places and the values shown in Table 4 are already rounded off.

Languages that have lower Dice's coefficient values than the threshold 0.60 are not considered closely-related languages. Based on the DCTP values, 74% of the domain languages were closely-related to each other. These languages were considered closely-related because they have more than 3 languages with DCTP values more than or equal to the threshold 0.60. The remaining 0.26% were considered outliers because these languages only have 1 closely-related language or no closely-related language at all. The DCTP values of the outliers ranges from 0.31-0.55. The outlier languages were Balangao and Bontok, Blaan, Chavacano, Ilocano, Iranun, Ivatan and Yami, Kallahan, Tiruray and Sama.

Comparing the DCTP values on the results using the farthest first algorithm, Cluster 0 have 7% difference including 2 outliers in the cluster and 93% closely-related languages. But this 93% cannot be concluded correctly clustered, these languages have to be clustered further, some languages were closely-related to it, but some are not and was only similar to its closely-related languages. Cluster 1 to 6 are clusters of languages that were considered outliers, and comparing it to the DCTP values, it is correct. All the languages included in Clusters 1, 2, 3, 4, 5 and 6 were considered outliers both in the results and DCTP evaluation. In Cluster 7, the DCTP values ranges from 0.65-0.76 validating that the languages in Cluster 7 were closely-related languages. Out of 8 clusters made for farthest first algorithm, only 1 cluster have some difference compared to its DCTP values, but this cluster has the most number of domain languages as its member with 30 languages. All in all, 56% of the domain languages were clustered differently based on its values and 44% of the domain languages were clustered in a manner that it's the same with its DCTP values using farthest first algorithm.

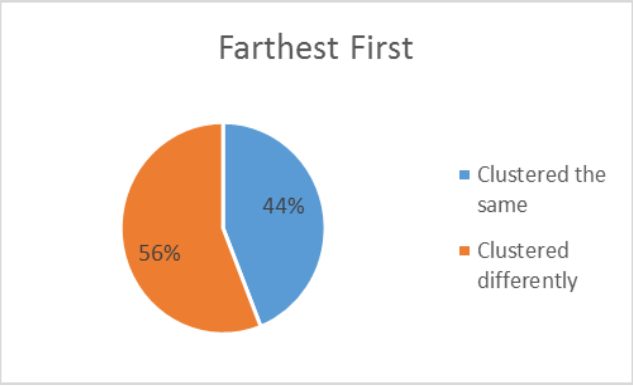

Figure 4-3. Farthest First Algorithm Evaluation using DCTP values

For simple k-means algorithm, 86% of the languages included in Cluster 0 were closely-related while the remaining 14% were outlier languages. In Cluster 1 and 2, 71% were closely-related languages while 29% of the languages included in the cluster were outliers. In Cluster 3, Buhid was not an outlier based on its DCTP value range that is 0.31-0.67 but was considered one when clustered using simple k-means. In Cluster 4, Balangao and Bontok were outliers, they are closely-related only to each other having a value of 0.65. In Cluster 5, Ifugao and Kallahan were grouped, Kallahan is an outlier language but not Ifugao, but they are closely-related to each other based on their values. Kallahan was considered outlier because it's only similar to Ifugao. In Cluster 6, all the languages included in this cluster were closely-related with values ranging from 0.64-0.67. In Cluster 7, 33% were closely-related, the other 66% were outliers and not closely-related. All in all, 74% of the domain languages were clustered the same while 26% were clustered differently.

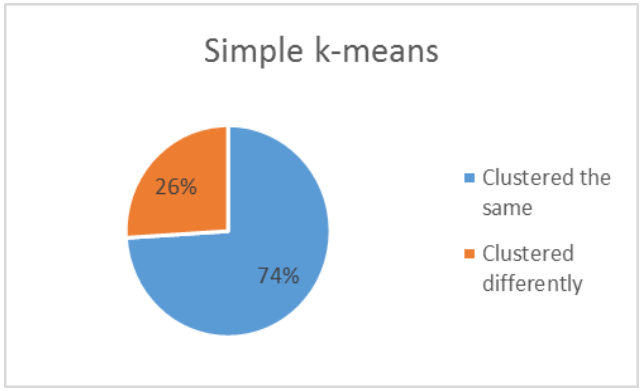

Figure 4-4. Result Evaluation using Simple k-means

The percentage used to show the similarity and difference between the results of clustering and its DCTP values came from the number of languages that are closely-related to each other and were grouped in the same cluster based on its values over the total number of domain languages.

$$Percentage = \frac{CRL}{TDL} \times 100$$

Equation 2. Percentage Formula

Aside from the metric used, the results were also compared to a Philippine Language Relations in a Map of Nathaniel Hermosa[4].

Based on the map, the smaller the node, the farther it is from the PMP (Proto-Malayo-Polynesian) branch, the bigger the node, the closer it is to the PMP branch. Comparison also depends on the thickness of the line that also depends on the node size that gives the relative closeness between the languages.

The languages included in the map were only 65%, 35% of the domain languages were not mentioned. Comparing the results using farthest first algorithm, 68% of the languages were clustered the same while 32% were clustered differently. Of these 32%, some the languages were not grouped with their similar languages, and some were supposed to be outlier languages. Using simple k-means algorithm, 61% of the languages were clustered the same and 39% were clustered differently.

[4]http://www.gmanetwork.com/news/story/331551/scitech/science/language-map-shows-philippine-languages-as-sibling-to-regional-tongues

The percentage used to compare the resulted language family tree with the language map was computed by counting the number of closely-related languages (based on the language map) and the number of languages in the same cluster (resulted clusters) that are found the same over the total number of domain languages.

The results were also compared to the Philippine Language Tree of Ethnologue. For the resulted clusters using farthest first algorithm 58% of the domain languages were closely-related based on the family tree and were grouped together with its similar languages. But still, these clusters should be clustered further because some of the languages were grouped with dissimilar languages. For simple k-means, 77% of the domain languages were grouped with its similar languages and 23% are not. Most of the languages were outlier languages that were included in a group of similar languages, and some dissimilar languages are still grouped with its similar languages.

5. CONCLUSIONS AND FURTHER WORK

In this paper, the results showed that the clusters made using trigram models presented group of languages that are similar and dissimilar. Languages that are dissimilar or not closely related to any of the domain languages are identified, some of the similar languages were grouped correctly based on the Dice's Coefficient value and matched the groups in the language family tree and language relational map while the other similar languages were grouped with both the closely-related and not.

Based from the three types of evaluation done with the results, there were three conclusions that can be drawn. First, with the use of character trigrams, similar languages can be identified and can be grouped in one cluster. Second, some dissimilar languages were joining similar languages in a group making a cluster bigger in number, because these dissimilar languages are joining the similar languages together with its similar languages instead of having its own cluster. The reason maybe the number of *k* used in the experiment. Third, there are outlier languages, languages that are alone in their cluster or having very few members in the cluster. These outlier languages sometimes were included in a cluster of similar languages.

Analyses showed that using character trigrams can identify closely-related languages. But considering only the orthography of the words in the languages are not enough to identify the language similarities. Results showed that the languages grouped in the same cluster may not be closely related and some are not in the same cluster but are closely related to each other.

It can be observed that the language clusters are composed of languages spoken in places located on the same region or is proximate to other places. It shows that languages may be influenced by their neighbor language and therefore, geographic location is also an important factor in language similarity. Phonetics as well is a significant aspect to consider in language similarity.

From the initial results, it has been concluded that some of the clusters made were not clearly grouped with the languages that are similar and dissimilar. Considering other features other than the trigram profiles of the languages can be used to cluster the languages more accurately.

For future work, the geographic location and phonetics will be explored as features; data where the domain languages are spoken per region and the phonetic alphabet will be collected. These data will be represented and will be processed to identify closely related languages. Furthermore, an updated language family tree of Philippine Languages will be constructed.

Philippine Language Family Tree – Outline (Ethnologue)

- I. Northern Philippines
  - A. Central Luzon
    - a. Ayta
    - b. Bolinao
    - c. Kapampangan
  - B. Northern Luzon
    - 1. Northern Cordillera
      - a. Agta
      - b. Itawit
      - c. Paranan
    - 2. Meso Cordillera
      - a. Balangao
      - b. Bontok
      - c. Ifugao
      - d. Kalinga
      - e. Kallahan
    - 3. Ilocano
    - 4. South Central Cordillera
      - a. Pangasinan
- II. Southern Philippines
  - A. Manobo
    - 1. North
      - a. Kagayanen
      - b. Binukid
    - 2. Central
      - a. Manobo
      - b. Matigsalug
  - B. Danao
    - 1. Maranao
    - 2. Iranun
    - 3. Maguindanao
  - C. Mindanao
    - 1. Chavacano
- III. Meso Philippines
  - A. Mangyan
    - 1. North
      - a. Iraya
      - b. Alangan
    - 2. South
      - a. Buhid
      - b. Hanunoo
  - B. Central Philippines
    - 1. Bikol
    - 2. Kalamian
      - a. Agutaynen
    - 3. Sambal
    - 4. Tagalog
    - 5. Mansaka
    - 6. Bilic
      - a. Tiruray
      - b. Blaan
    - 7. Basliic
      - a. Ivatan
    - 8. Bisaya
      - a. Central
        - Cebuano
        - Hiligaynon
        - Masbatenyo
        - Waray
      - b. West
        - Kiniray-a
      - c. South
        - Tausug
- IV. Sama Badjaw
  - A. Inabaknon
  - B. Sama
  - C. Yakan

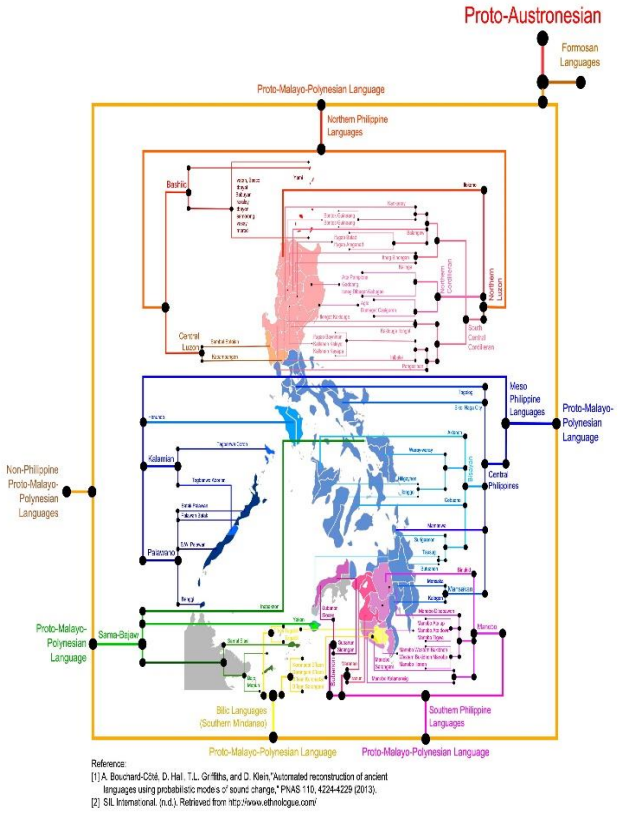

Figure 5-1. Philippine Language Relations in a Map[5]

6. ACKNOWLEDGEMENT

Our thanks to the following: Judi Diane Miñon, Manolito Octaviano Jr., Mr. Leif Romeritch Syliongka, Dana Genevieve Macabante, John Casper Tambanillo and Nove Ellema for being instrumental to this research.

7. REFERENCES

[1] Enright, J. and Kondrak, G.: The application of chordal graphs to inferring phylogenetic trees of languages. In: Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 545–552, Chiang Mai, Thailand (2011).

[2] Octaviano Jr., M., Fajutagana, R., Lim, C.M., Miñon, J.D., Morano, J., Tinoco, R.C., Oco, N.: The use of Trigram Models in Classifying and Clustering different Philippine Languages. In: The Tenth International Conference on Knowledge, Information and Creativity Support Systems (KICSS), pp. 546-552. (2015)

[3] Oco, N., Syliongka, L.R., Roxas, R.E.: Clustering Philippine Languages. In: 16th Philippine Computing Science Congress. (2016)

[4] Oco, N., Sison-Buban, R., Syliongka, L.R., Roxas, R.E., Ilao, J.: Trigram Ranking: Metric for Language Similarity and Clustering. Malay, pp. 53-68 (2014)

[5] Oco, N., Syliongka, L.R., Roxas, R.E., Ilao, J.: Dice's Coefficient on Trigram Profiles as Metric for Language Similarity. In: Oriental chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (OCOCOSDA). (2013)

[6] Octaviano Jr., M., Hiya, A.P., Oco, N.: Implementing a Trigram-based Language Identification System for Philippine Languages. In: 11th National Natural Language Processing Research Symposium, pp. 38–42. Computing Society of the Philippines – Special Interest Group on Natural Language Processing, Manila (2015)

[7] Rama, T. and Singh, A.K.: From Bag of Languages to Family Trees from Noisy Corpus. In: International Conference RANLP 2009 - Borovets, Bulgaria, pages 355–359. (2009)

[8] Nagata, R. and Whittaker, E.: Reconstructing an Indo-European Family Tree from Non-native English texts. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1137–1147, Sofia, Bulgaria. Association for Computational Linguistics (2013)

[9] Kitchen, A., Ehret, C., Assefa, S., Mulligan, C.: Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. In: Proceedings of the Royal Society. (2009)

[10] Chang, W. and Cathcart, C.: Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. (2015)

[11] Xu, R. and Wunsch, D. 2009. Clustering. IEEE Press Series on Computational Intelligence.

[12] Biber, D. 1993. 'Representativeness in corpus design'. Literary and Linguistic Computing 8/4: 243-57.

[13] Lüdeling, Anke (2006) Using corpora in the classification of language relationships. In: Zeitschrift für Anglistik und Amerikanistik. Special Issue on 'The Scope and Limits of Corpus Linguistics' (guest editor: Volker Gast), 217-227.

[14] Sarmah, Das, Gogi, and Horo L. (2012), Phylogenetic Analysis of a few Languages of Assam

[5] https://imphscience.wordpress.com/2013/10/15/philippine-language-relations-in-a-map/

8. APPENDIX

Figure 8-1. Philippine Language Family Tree (Farthest First Algorithm)

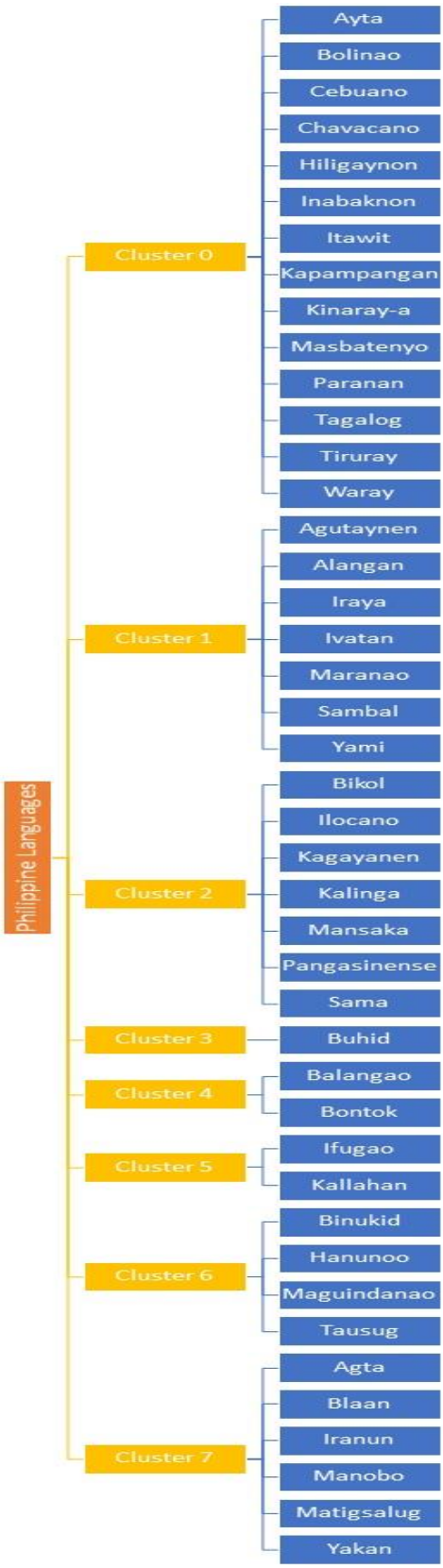

Figure 8-2. Philippine Language Family Tree (Simple k-means Algorithm)

Table 3. Trigram Profiles of Domain Languanges

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
_na	ng_	_ka	an_	_he	an_	an_	_i_	an_	an_	ang	ng_	el_	_ka	ng	an	ti_	_si	_ba	_da	an_	ka	_na	an_	an_	ng	ang	ka	_sa	ang	an_	an_	te_	an_	en_	ta_	ay_	ng	in_	_i_	an_	_si	ma
na_	ang	_in	_ya	hen	_na	_ha	_di	_ni	_na	an_	ang	_el	an_	_sa	_na	_ke	si_	shi	_sa	nga	an_	na_	_ma	_ni	ing	_ka	an_	sa_	ng_	_a_	_na	_te	ay_	_a_	ke_	_ha	ang	an_	_m	nga	in_	an_
ta_	_ta	in_	ya_	en_	_sa	_hu	_na	in_	nan	_ha	sa_	_co	_ma	ang	_di	iti	an_	na_	da_	_ka	_na	_ta	_ka	ni_	ang	ng_	_na	_di	an_	iya	_sa	ne_	_na	ay_	_no	ya_	_sa	_ma	b_	_ng	si_	_o_
_ta	tan	ang	_na	ay_	_ka	_sa	di_	_sa	ay_	_ka	_sa	_ma	_sa	ga_	ya_	_a	_na	_mu	ko	_ya	ma	an_	un_	_et	an_	ga_	na_	di_	_na	_ma	_ka	_ne	_si	an_	ho_	an_	sa_	_si	de_	_ha	an_	ang
ya_	_an	ay_	ay_	_an	ng_	ha_	an_	nin	cha	_ma	_ka	ya_	sa_	_ka	nan	_ti	ga_	_na	an_	_ng	a_	ta_	_na	_ka	ka	_an	nu_	ya_	_ya	ang	na_	an_	_ta	di_	ma_	_ya	_an	iya	_de	ga_	_ma	ka
_a_	ong	an_	_ka	cha	_an	hu_	na_	_na	_ay	fa	an_	_ya	ang	sa_	_ya	et_	na_	la_	_ba	ya_	no_	_pa	_si	_tu	na	nga	_ma	_eg	_ka	na_	_ma	_ka	_sa	_na	oi_	in_	_na	in	_do	_hi	_pa	_ya
an_	_on	_sa	_ha	an_	_ni	aw_	_da	_ma	as_	nga	ga_	con	ay_	_na	ana	na_	_pa	ba_	ba_	kan	na_	en_	_da	ey_	at_	_na	_a_	iya	_ma	na_	an	ka_	_ma	di	mi_	_ni	at_	ang	_n_	_an	_in	ng_
_ya	an_	sa_	_n_	_wa	ang	_da	ad_	na_	_as	ya_	_an	on_	man	nga	_da	ket	_i_	ti_	sa_	ga_	do_	_ma	di	_hi	_in	_ng	kan	da_	yan	man	sa_	_eg	_to	_te	_hi	_hi	_ka	_ka	ob_	han	_du	na_
ang	_na	kan	_hi	chi	_si	sa_	dad	_a_	_ch	_na	nga	ta_	kan	_an	_an	da_	nga	aba	_na	_na	do	_da	na_	et_	_ma	_sa	_sa	en	na_	_ka	san	ey_	en_	_en	_dz	_ma	_ng	na	dob	_na	ne_	ya_
en_	_ma	_ag	ang	way	_ma	ta	_ma	_ko	_ka	ga_	_ma	_de	_pa	_si	_hi	_da	_ka	li_	_ka	na_	no	aw_	_un	da_	na_	kan	_ni	_da	aw_	_so	_pa	_si	_ka	_ma	hii	_na	_at	sin	_t_	_ka	ang	en_

Table 4. Dice's Coefficient Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	12	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
1	1.0	0.5	0.5	0.6	0.5	0.6	0.6	0.5	0.6	0.5	0.5	0.6	0.4	0.6	0.6	0.6	0.5	0.6	0.4	0.5	0.7	0.5	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.5	0.6	0.6	0.6	0.6	0.4	0.5	0.6	0.6	0.5	0.6	0.6	0.4
2	0.5	1.0	0.7	0.6	0.5	0.6	0.6	0.4	0.6	0.5	0.6	0.6	0.5	0.4	0.6	0.5	0.6	0.6	0.4	0.7	0.6	0.6	0.6	0.6	0.5	0.6	0.6	0.5	0.5	0.6	0.7	0.6	0.5	0.7	0.6	0.4	0.6	0.6	0.4	0.5	0.6	0.5	0.5
3	0.5	0.7	1.0	0.5	0.4	0.6	0.5	0.4	0.6	0.4	0.6	0.5	0.4	0.5	0.5	0.5	0.5	0.5	0.4	0.8	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.5	0.5	0.6	0.7	0.5	0.5	0.6	0.5	0.4	0.7	0.6	0.4	0.5	0.5	0.5	0.5
4	0.6	0.6	0.5	1.0	0.5	0.7	0.6	0.4	0.6	0.5	0.5	0.7	0.4	0.6	0.7	0.6	0.5	0.7	0.4	0.6	0.6	0.5	0.6	0.6	0.5	0.6	0.7	0.9	0.5	0.6	0.5	0.7	0.5	0.6	0.6	0.4	0.6	0.6	0.5	0.5	0.7	0.5	0.4
5	0.5	0.5	0.4	0.5	1.0	0.5	0.5	0.4	0.5	0.6	0.4	0.5	0.4	0.4	0.5	0.5	0.5	0.5	0.4	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.4	0.4	0.5	0.5	0.5	0.4	0.5	0.5	0.3	0.5	0.5	0.4	0.5	0.5	0.5	0.4
6	0.6	0.6	0.6	0.7	0.5	1.0	0.6	0.5	0.6	0.5	0.6	0.7	0.5	0.6	0.7	0.6	0.6	0.7	0.5	0.6	0.6	0.5	0.6	0.6	0.5	0.6	0.7	0.5	0.5	0.7	0.6	0.8	0.5	0.7	0.7	0.4	0.6	0.7	0.6	0.5	0.7	0.6	0.4
7	0.6	0.6	0.5	0.6	0.5	0.6	1.0	0.5	0.6	0.5	0.4	0.7	0.4	0.7	0.7	0.6	0.5	0.6	0.4	0.5	0.6	0.5	0.6	0.6	0.6	0.7	0.7	0.7	0.6	0.6	0.5	0.6	0.7	0.6	0.7	0.4	0.5	0.6	0.6	0.6	0.6	0.6	0.4
8	0.5	0.4	0.4	0.4	0.4	0.5	0.5	1.0	0.4	0.4	0.5	0.3	0.5	0.5	0.5	0.4	0.5	0.4	0.4	0.4	0.4	0.5	0.5	0.4	0.5	0.5	0.5	0.5	0.4	0.4	0.4	0.5	0.5	0.5	0.3	0.4	0.5	0.4	0.5	0.5	0.5	0.4	0.4
9	0.6	0.6	0.6	0.6	0.5	0.6	0.6	0.4	1.0	0.5	0.5	0.6	0.4	0.6	0.6	0.6	0.6	0.7	0.5	0.6	0.6	0.5	0.7	0.6	0.5	0.6	0.6	0.6	0.5	0.6	0.6	0.7	0.5	0.7	0.6	0.4	0.6	0.6	0.5	0.5	0.6	0.5	0.4
10	0.5	0.5	0.4	0.5	0.6	0.5	0.5	0.4	0.5	1.0	0.4	0.5	0.4	0.5	0.5	0.5	0.5	0.5	0.4	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.4	0.4	0.5	0.5	0.5	0.5	0.5	0.4
11	0.5	0.6	0.6	0.5	0.4	0.6	0.4	0.5	0.5	0.4	1.0	0.5	0.3	0.5	0.5	0.5	0.4	0.5	0.4	0.7	0.4	0.5	0.5	0.5	0.3	0.4	0.5	0.4	0.4	0.6	0.6	0.5	0.4	0.5	0.5	0.3	0.6	0.5	0.4	0.5	0.6	0.4	0.4
12	0.6	0.6	0.5	0.7	0.5	0.7	0.7	0.3	0.6	0.5	0.5	1.0	0.4	0.6	0.8	0.6	0.6	0.7	0.5	0.6	0.6	0.5	0.6	0.7	0.5	0.6	0.8	0.6	0.5	0.7	0.6	0.7	0.5	0.6	0.6	0.4	0.6	0.7	0.6	0.5	0.8	0.6	0.4
12	0.4	0.5	0.4	0.4	0.4	0.5	0.4	0.5	0.4	0.4	0.3	0.4	1.0	0.3	0.4	0.4	0.5	0.4	0.4	0.4	0.5	0.5	0.5	0.4	0.4	0.5	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.5	0.4	0.3	0.4	0.4	0.3	0.4	0.4	0.4	0.4
14	0.6	0.4	0.5	0.6	0.4	0.6	0.7	0.5	0.6	0.5	0.5	0.6	0.3	1.0	0.6	0.5	0.5	0.7	0.4	0.5	0.5	0.4	0.6	0.6	0.5	0.6	0.7	0.6	0.5	0.6	0.4	0.6	0.6	0.5	0.6	0.3	0.5	0.6	0.7	0.5	0.6	0.6	0.3
15	0.6	0.6	0.5	0.7	0.5	0.7	0.7	0.5	0.6	0.5	0.5	0.8	0.4	0.6	1.0	0.6	0.6	0.7	0.5	0.6	0.6	0.5	0.7	0.6	0.5	0.6	0.8	0.5	0.5	0.6	0.5	0.8	0.5	0.6	0.6	0.4	0.6	0.7	0.6	0.5	0.8	0.6	0.4
16	0.6	0.5	0.5	0.6	0.5	0.6	0.6	0.4	0.6	0.5	0.5	0.6	0.4	0.5	0.6	1.0	0.5	0.6	0.4	0.5	0.6	0.5	0.5	0.6	0.6	0.6	0.6	0.5	0.5	0.5	0.5	0.6	0.5	0.6	0.6	0.3	0.5	0.6	0.5	0.5	0.6	0.6	0.4
17	0.5	0.6	0.5	0.5	0.5	0.6	0.5	0.5	0.6	0.5	0.4	0.6	0.5	0.5	0.6	0.5	1.0	0.6	0.4	0.5	0.5	0.5	0.6	0.6	0.5	0.6	0.6	0.5	0.5	0.5	0.5	0.6	0.5	0.6	0.6	0.3	0.5	0.6	0.5	0.5	0.6	0.5	0.4
18	0.6	0.6	0.5	0.7	0.5	0.7	0.6	0.4	0.7	0.5	0.5	0.7	0.4	0.7	0.7	0.6	0.6	1.0	0.4	0.6	0.6	0.5	0.6	0.6	0.5	0.6	0.7	0.6	0.5	0.6	0.5	0.7	0.6	0.6	0.7	0.4	0.6	0.7	0.6	0.5	0.7	0.6	0.4
19	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.5	0.4	0.4	0.5	0.4	0.4	0.5	0.4	0.4	0.4	1.0	0.4	0.4	0.4	0.5	0.5	0.4	0.5	0.4	0.5	0.5	0.4	0.4	0.4	0.5	0.5	0.5	0.4	0.4	0.5	0.4	0.4	0.4	0.5	0.3
20	0.5	0.7	0.8	0.6	0.4	0.6	0.5	0.4	0.6	0.4	0.7	0.6	0.4	0.5	0.6	0.5	0.5	0.6	0.4	1.0	0.5	0.6	0.6	0.5	0.4	0.5	0.6	0.5	0.5	0.6	0.7	0.6	0.5	0.6	0.5	0.4	0.7	0.6	0.5	0.5	0.6	0.5	0.5
21	0.7	0.6	0.5	0.6	0.5	0.6	0.6	0.4	0.6	0.5	0.4	0.6	0.5	0.5	0.6	0.6	0.5	0.6	0.4	0.5	1.0	0.5	0.6	0.6	0.6	0.6	0.6	0.6	0.5	0.6	0.5	0.6	0.5	0.6	0.6	0.4	0.5	0.5	0.5	0.5	0.6	0.6	0.4
22	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.5	0.5	0.4	0.6	0.5	1.0	0.5	0.5	0.4	0.5	0.5	0.4	0.4	0.5	0.5	0.5	0.4	0.6	0.5	0.3	0.6	0.5	0.3	0.4	0.5	0.4	0.5
23	0.6	0.6	0.5	0.6	0.5	0.6	0.6	0.5	0.7	0.5	0.5	0.6	0.5	0.6	0.7	0.5	0.6	0.6	0.5	0.6	0.6	0.5	1.0	0.6	0.5	0.6	0.7	0.6	0.5	0.6	0.5	0.7	0.5	0.6	0.6	0.4	0.5	0.6	0.5	0.5	0.6	0.5	0.4
24	0.6	0.6	0.5	0.6	0.5	0.6	0.6	0.4	0.6	0.5	0.5	0.7	0.4	0.6	0.6	0.6	0.6	0.6	0.5	0.5	0.6	0.5	0.6	1.0	0.5	0.6	0.6	0.6	0.5	0.7	0.6	0.6	0.5	0.6	0.7	0.4	0.5	0.6	0.6	0.5	0.6	0.6	0.4
25	0.6	0.5	0.4	0.5	0.5	0.5	0.6	0.5	0.5	0.5	0.3	0.5	0.4	0.5	0.5	0.6	0.5	0.5	0.4	0.6	0.4	0.6	0.4	0.5	0.5	1.0	0.6	0.5	0.6	0.6	0.4	0.5	0.5	0.6	0.5	0.6	0.3	0.4	0.5	0.5	0.5	0.6	0.3
26	0.6	0.6	0.5	0.6	0.5	0.6	0.7	0.5	0.6	0.5	0.4	0.6	0.5	0.6	0.6	0.6	0.6	0.6	0.5	0.5	0.6	0.5	0.6	0.6	0.6	1.0	0.6	0.7	0.6	0.6	0.5	0.6	0.6	0.6	0.7	0.4	0.5	0.6	0.6	0.5	0.6	0.6	0.4
27	0.6	0.6	0.5	0.7	0.5	0.7	0.7	0.5	0.6	0.5	0.5	0.8	0.4	0.7	0.8	0.6	0.6	0.7	0.4	0.6	0.6	0.5	0.7	0.6	0.5	0.6	1.0	0.6	0.5	0.7	0.5	0.8	0.6	0.6	0.7	0.4	0.5	0.7	0.6	0.5	0.8	0.6	0.4
28	0.6	0.5	0.5	0.9	0.4	0.5	0.7	0.5	0.6	0.5	0.4	0.6	0.4	0.6	0.5	0.5	0.5	0.6	0.5	0.5	0.6	0.4	0.6	0.6	0.6	0.7	0.6	1.0	0.6	0.5	0.6	0.5	0.6	0.6	0.6	0.4	0.4	0.6	0.7	0.5	0.7	0.4	0.6
29	0.6	0.5	0.5	0.5	0.4	0.5	0.6	0.4	0.5	0.4	0.4	0.5	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.6	0.6	0.5	0.6	1.0	0.5	0.5	0.5	0.6	0.5	0.6	0.4	0.5	0.6	0.5	0.5	0.5	0.6	0.4
30	0.6	0.6	0.6	0.6	0.5	0.7	0.6	0.4	0.6	0.5	0.6	0.7	0.4	0.6	0.6	0.5	0.5	0.6	0.4	0.6	0.6	0.5	0.6	0.7	0.4	0.6	0.7	0.5	0.5	1.0	0.6	0.6	0.5	0.6	0.6	0.4	0.6	0.6	0.5	0.7	0.5	0.4	0.4
31	0.5	0.7	0.7	0.5	0.5	0.6	0.5	0.4	0.6	0.5	0.6	0.6	0.4	0.4	0.5	0.5	0.5	0.5	0.4	0.7	0.5	0.5	0.5	0.0	0.5	0.5	0.5	0.6	0.5	0.6	1.0	0.5	0.5	0.6	0.5	0.5	0.7	0.6	0.4	0.5	0.6	0.5	0.5
32	0.6	0.6	0.5	0.7	0.5	0.8	0.6	0.5	0.7	0.5	0.5	0.7	0.5	0.6	0.8	0.6	0.6	0.7	0.4	0.6	0.6	0.5	0.7	0.6	0.5	0.6	0.8	0.5	0.5	0.6	0.5	1.0	0.5	0.6	0.7	0.4	0.5	0.7	0.6	0.5	0.8	0.5	0.4
33	0.6	0.5	0.5	0.5	0.4	0.5	0.7	0.5	0.5	0.5	0.4	0.5	0.4	0.6	0.5	0.5	0.5	0.6	0.5	0.5	0.5	0.4	0.5	0.5	0.6	0.6	0.6	0.6	0.6	0.5	0.5	0.5	1.0	0.5	0.6	0.4	0.4	0.6	0.6	0.5	0.5	0.6	0.4
34	0.6	0.7	0.6	0.6	0.5	0.7	0.6	0.5	0.7	0.5	0.5	0.6	0.5	0.5	0.6	0.6	0.6	0.6	0.5	0.6	0.6	0.6	0.6	0.6	0.5	0.6	0.6	0.6	0.5	0.6	0.6	0.6	0.5	1.0	0.6	0.4	0.6	0.6	0.5	0.5	0.6	0.5	0.5
35	0.6	0.6	0.5	0.6	0.5	0.7	0.7	0.3	0.6	0.5	0.5	0.6	0.4	0.6	0.6	0.6	0.6	0.7	0.5	0.5	0.6	0.5	0.6	0.7	0.6	0.7	0.7	0.6	0.6	0.6	0.5	0.7	0.6	0.6	1.0	0.4	0.5	0.6	0.6	0.5	0.6	0.6	0.4
36	0.4	0.4	0.4	0.4	0.3	0.4	0.4	0.4	0.4	0.4	0.3	0.4	0.3	0.3	0.4	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.3	0.4	0.4	0.3	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.4	1.0	0.4	0.4	0.3	0.4	0.4	0.4	0.4
37	0.5	0.6	0.7	0.6	0.5	0.6	0.5	0.5	0.6	0.4	0.6	0.6	0.4	0.5	0.6	0.5	0.5	0.6	0.4	0.7	0.5	0.6	0.5	0.5	0.4	0.5	0.5	0.4	0.5	0.6	0.7	0.5	0.4	0.6	0.5	0.4	1.0	0.6	0.4	0.5	0.6	0.5	0.4
38	0.6	0.6	0.6	0.6	0.5	0.7	0.6	0.4	0.6	0.5	0.5	0.7	0.4	0.6	0.7	0.6	0.6	0.7	0.5	0.6	0.5	0.5	0.6	0.6	0.5	0.6	0.7	0.6	0.6	0.6	0.6	0.7	0.6	0.6	0.6	0.4	0.6	1.0	0.6	0.5	0.7	0.6	0.4
39	0.6	0.4	0.4	0.5	0.4	0.6	0.6	0.5	0.5	0.5	0.4	0.6	0.3	0.7	0.6	0.5	0.5	0.6	0.4	0.5	0.5	0.3	0.5	0.6	0.5	0.6	0.6	0.7	0.5	0.6	0.4	0.6	0.6	0.5	0.6	0.3	0.4	0.6	1.0	0.5	0.6	0.5	0.6
40	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.5	1.0	0.5	0.5	0.4
41	0.6	0.6	0.5	0.7	0.5	0.7	0.6	0.5	0.6	0.5	0.6	0.8	0.4	0.6	0.8	0.6	0.6	0.7	0.4	0.6	0.6	0.5	0.6	0.6	0.5	0.6	0.8	0.7	0.5	0.7	0.6	0.8	0.5	0.6	0.6	0.4	0.6	0.7	0.6	0.5	1.0	0.5	0.4
42	0.6	0.5	0.5	0.5	0.5	0.6	0.6	0.4	0.5	0.5	0.4	0.6	0.4	0.6	0.6	0.6	0.5	0.6	0.5	0.5	0.6	0.4	0.5	0.6	0.6	0.6	0.6	0.4	0.6	0.5	0.5	0.5	0.6	0.5	0.6	0.4	0.5	0.6	0.6	0.5	0.5	1.0	0.4
43	0.4	0.5	0.5	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.3	0.4	0.4	0.4	0.4	0.3	0.5	0.4	0.5	0.4	0.4	0.3	0.4	0.4	0.6	0.4	0.4	0.5	0.4	0.4	0.5	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	1.0