Online ISSN : 2249-460X
Print ISSN : 0975-587X
DOI : 10.17406/GJHSS

Global Journal

of Human Social Sciences: G

Linguistics & Education

Highlights

A Sociolinguistic Study

Role of a Teacher in Teaching

Coefficients and Creation

Normalization and Dimensionality

Discovering Thoughts, Inventing Future

Volume 16 ISSUE 3 VERSION 1.0

© 2001-2016 by Global Journal of Human Social Sciences, USA

Volume 16 Issue 12 was originally issued as G3

Global Journal of Human-Social Science: G
Linguistics & Education

Volume 16 Issue 12
Originally issued as G3

Global Journal of Human-Social Science: G

Linguistics & Education

Volume 16 Issue 3 (Ver. 1.0)

Open Association of Research Society

© Global Journal of Human Social Sciences. 2016.

All rights reserved.

This is a special issue published in version 1.0 of "Global Journal of Human Social Sciences." By Global Journals Inc.

All articles are open access articles distributed under "Global Journal of Human Social Sciences"

Reading License, which permits restricted use. Entire contents are copyright by of "Global Journal of Human Social Sciences" unless otherwise noted on specific articles.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without written permission.

The opinions and statements made in this book are those of the authors concerned. Ultraculture has not verified and neither confirms nor denies any of the foregoing and no warranty or fitness is implied.

Engage with the contents herein at your own risk.

The use of this journal, and the terms and conditions for our providing information, is governed by our Disclaimer, Terms and Conditions and Privacy Policy given on our website http://globaljournals.us/terms-and-condition/menu-id-1463/

By referring / using / reading / any type of association / referencing this journal, this signifies and you acknowledge that you have read them and that you accept and will be bound by the terms thereof.

All information, journals, this journal, activities undertaken, materials, services and our website, terms and conditions, privacy policy, and this journal is subject to change anytime without any prior notice.

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

*(A Delaware USA Incorporation with "Good Standing"; **Reg. Number: 0423089**)*

Sponsors: Open Association of Research Society
Open Scientific Standards

Publisher's Headquarters office

Global Journals® Headquarters
945th Concord Streets,
Framingham Massachusetts Pin: 01701,
United States of America
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Journals Incorporated
2nd, Lansdowne, Lansdowne Rd., Croydon-Surrey,
Pin: CR9 2ER, United Kingdom

Packaging & Continental Dispatching

Global Journals
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin:452009, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investors@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Including by Air Parcel Charges):

For Authors:
22 USD (B/W) & 50 USD (Color)
Yearly Subscription (Personal & Institutional):
200 USD (B/W) & 250 USD (Color)

Integrated Editorial Board
(Computer Science, Engineering, Medical, Management, Natural Science, Social Science)

John A. Hamilton,"Drew" Jr.,
Ph.D., Professor, Management
Computer Science and Software Engineering
Director, Information Assurance Laboratory
Auburn University

Dr. Henry Hexmoor
IEEE senior member since 2004
Ph.D. Computer Science, University at Buffalo
Department of Computer Science
Southern Illinois University at Carbondale

Dr. Osman Balci, Professor
Department of Computer Science
Virginia Tech, Virginia University
Ph.D. and M. S. Syracuse University, Syracuse, New York
M. S. and B. S. Bogazici University, Istanbul, Turkey

Yogita Bajpai
M.Sc. (Computer Science), FICCT
U.S.A.Email:
yogita@computerresearch.org

Dr. T. David A. Forbes
Associate Professor and Range Nutritionist
Ph.D. Edinburgh University - Animal Nutrition
M. S. Aberdeen University - Animal Nutrition
B. A. University of Dublin- Zoology

Dr. Wenying Feng
Professor, Department of Computing & Information Systems
Department of Mathematics
Trent University, Peterborough,
ON Canada K9J 7B8

Dr. Thomas Wischgoll
Computer Science and Engineering,
Wright State University, Dayton, Ohio
B.S., M.S., Ph.D.
(University of Kaiserslautern)

Dr. Abdurrahman Arslanyilmaz
Computer Science & Information Systems Department
Youngstown State University
Ph.D., Texas A&M University
University of Missouri, Columbia
Gazi University, Turkey

Dr. Xiaohong He
Professor of International Business
University of Quinnipiac
BS, Jilin Institute of Technology; MA, MS, PhD,. (University of Texas-Dallas)

Burcin Becerik-Gerber
University of Southern California
Ph.D. in Civil Engineering
DDes from Harvard University
M. S. from University of California, Berkeley
& Istanbul University

Dr. Bart Lambrecht
Director of Research in Accounting and Finance Professor of Finance
Lancaster University Management School
BA (Antwerp); MPhil, MA, PhD (Cambridge)

Dr. Carlos García Pont
Associate Professor of Marketing
IESE Business School, University of Navarra
Doctor of Philosophy (Management), Massachusetts Institute of Technology (MIT)
Master in Business Administration, IESE, University of Navarra
Degree in Industrial Engineering, Universitat Politècnica de Catalunya

Dr. Fotini Labropulu
Mathematics - Luther College
University of ReginaPh.D., M.Sc. in Mathematics
B.A. (Honors) in Mathematics
University of Windso

Dr. Lynn Lim
Reader in Business and Marketing
Roehampton University, London
BCom, PGDip, MBA (Distinction), PhD, FHEA

Dr. Mihaly Mezei
ASSOCIATE PROFESSOR
Department of Structural and Chemical Biology, Mount Sinai School of Medical Center
Ph.D., Etvs Lornd University
Postdoctoral Training,
New York University

Dr. Söhnke M. Bartram
Department of Accounting and Finance Lancaster University Management SchoolPh.D. (WHU Koblenz)
MBA/BBA (University of Saarbrücken)

Dr. Miguel Angel Ariño
Professor of Decision Sciences
IESE Business School
Barcelona, Spain (Universidad de Navarra)
CEIBS (China Europe International Business School).
Beijing, Shanghai and Shenzhen
Ph.D. in Mathematics
University of Barcelona
BA in Mathematics (Licenciatura)
University of Barcelona

Philip G. Moscoso
Technology and Operations Management
IESE Business School, University of Navarra
Ph.D in Industrial Engineering and Management, ETH Zurich
M.Sc. in Chemical Engineering, ETH Zurich

Dr. Sanjay Dixit, M.D.
Director, EP Laboratories, Philadelphia VA Medical Center
Cardiovascular Medicine - Cardiac Arrhythmia
Univ of Penn School of Medicine

Dr. Han-Xiang Deng
MD., Ph.D
Associate Professor and Research
Department Division of Neuromuscular Medicine
Davee Department of Neurology and Clinical Neuroscience Northwestern University
Feinberg School of Medicine

Dr. Pina C. Sanelli
Associate Professor of Public Health
Weill Cornell Medical College
Associate Attending Radiologist
NewYork-Presbyterian Hospital
MRI, MRA, CT, and CTA
Neuroradiology and Diagnostic Radiology
M.D., State University of New York at Buffalo, School of Medicine and Biomedical Sciences

Dr. Roberto Sanchez
Associate Professor
Department of Structural and Chemical Biology
Mount Sinai School of Medicine
Ph.D., The Rockefeller University

Dr. Wen-Yih Sun
Professor of Earth and Atmospheric Sciences Purdue University Director National Center for Typhoon and Flooding Research, Taiwan
University Chair Professor
Department of Atmospheric Sciences, National Central University, Chung-Li, Taiwan University Chair Professor
Institute of Environmental Engineering, National Chiao Tung University, Hsin-chu, Taiwan. Ph.D., MS The University of Chicago, Geophysical Sciences
BS National Taiwan University, Atmospheric Sciences
Associate Professor of Radiology

Dr. Michael R. Rudnick
M.D., FACP
Associate Professor of Medicine
Chief, Renal Electrolyte and Hypertension Division (PMC)
Penn Medicine, University of Pennsylvania
Presbyterian Medical Center, Philadelphia
Nephrology and Internal Medicine
Certified by the American Board of Internal Medicine

Dr. Bassey Benjamin Esu
B.Sc. Marketing; MBA Marketing; Ph.D Marketing
Lecturer, Department of Marketing, University of Calabar
Tourism Consultant, Cross River State Tourism Development Department
Co-ordinator , Sustainable Tourism Initiative, Calabar, Nigeria

Dr. Aziz M. Barbar, Ph.D.
IEEE Senior Member
Chairperson, Department of Computer Science
AUST - American University of Science & Technology
Alfred Naccash Avenue – Ashrafieh

President Editor (HON.)

Dr. George Perry, (Neuroscientist)
Dean and Professor, College of Sciences
Denham Harman Research Award (American Aging Association)
ISI Highly Cited Researcher, Iberoamerican Molecular Biology Organization
AAAS Fellow, Correspondent Member of Spanish Royal Academy of Sciences
University of Texas at San Antonio
Postdoctoral Fellow (Department of Cell Biology)
Baylor College of Medicine
Houston, Texas, United States

Chief Author (HON.)

Dr. R.K. Dixit
M.Sc., Ph.D., FICCT
Chief Author, India
Email: authorind@computerresearch.org

Dean & Editor-in-Chief (HON.)

Vivek Dubey(HON.)
MS (Industrial Engineering),
MS (Mechanical Engineering)
University of Wisconsin, FICCT
Editor-in-Chief, USA
editorusa@computerresearch.org

Sangita Dixit
M.Sc., FICCT
Dean & Chancellor (Asia Pacific)
deanind@computerresearch.org

Suyash Dixit
(B.E., Computer Science Engineering), FICCTT
President, Web Administration and
Development , CEO at IOSRD
COO at GAOR & OSS

Er. Suyog Dixit
(M. Tech), BE (HONS. in CSE), FICCT
SAP Certified Consultant
CEO at IOSRD, GAOR & OSS
Technical Dean, Global Journals Inc. (US)
Website: www.suyogdixit.com
Email:suyog@suyogdixit.com

Pritesh Rajvaidya
(MS) Computer Science Department
California State University
BE (Computer Science), FICCT
Technical Dean, USA
Email: pritesh@computerresearch.org

Luis Galárraga
J!Research Project Leader
Saarbrücken, Germany

Contents of the Issue

i. Copyright Notice
ii. Editorial Board Members
iii. Chief Author and Dean
iv. Contents of the Issue

1. The Implications of Girl-Child Education to Nation Building in the 21st Century in Nigeria. *1-4*
2. Trends and Prospects of Instructional Material Development and Delivery at the University of Zambia. *5-11*
3. A Sociolinguistic Study of Choosing Names for Newborn Children in Jordan. *13-22*
4. Agglomerative Hierarchical Clustering: An Introduction to Essentials. (1) Proximity Coefficients and Creation of a Vector-Distance Matrix and (2) Construction of the Hierarchical Tree and a Selection of Methods. *23-50*
5. Role of a Teacher in Teaching Speaking by Following a Communicative Approach: To What Extent is this Possible in an ESL Context like Bangladesh? *51-54*
6. Agglomerative Hierarchical Clustering: An Introduction to Essentials. (3) Standardization, Normalization, and Dimensionality Reduction of a Data Matrix. *55-63*

v. Fellows
vi. Auxiliary Memberships
vii. Process of Submission of Research Paper
viii. Preferred Author Guidelines
ix. Index

Global Journal of human-social science: G
Linguistics & Education
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-460X & Print ISSN: 0975-587X

The Implications of Girl-Child Education to Nation Building in the 21st Century in Nigeria

By Omede Andrew A. & Agahiu Grace Etumabo

Kogi State College of Education, Nigeria

Abstract- This paper examines the implications of Girl-child to nation building in the 21st century in Nigeria. The paper began by pointing out the wrong notions that many Nigerians have particularly the rural dwellers about women being consider as properties for man and objects for their pleasure and how this notion restrains them from training their girl-children in schools. The paper further examined the concept of girl-child education to be all inclusive, some hindrances to effective girl-child education such as economic factors, sexual violence and abuse, political factors, the school environmental factors and socio-cultural and religious factors were highlighted. Included in the paper also was the implications that effective girl-child education would have on nation building such as poverty-reducing effects, improves health and nutrition, reduces inequality, reduces women's fertility rates, lowers infant and mortality rates and increases women's labour force participation rates and earnings.

Keywords: girl-child, education, implication, nation, building, 21st century Nigeria.

GJHSS-G Classification : FOR Code: 139999

Strictly as per the compliance and regulations of:

© 2016. Omede Andrew A. & Agahiu Grace Etumabo. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http:// creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Implications of Girl-Child Education to Nation Building in the 21st Century in Nigeria

Omede Andrew A. [α] & Agahiu Grace Etumabo [σ]

Abstract- This paper examines the implications of Girl-child to nation building in the 21st century in Nigeria. The paper began by pointing out the wrong notions that many Nigerians have particularly the rural dwellers about women being consider as properties for man and objects for their pleasure and how this notion restrains them from training their girl-children in schools. The paper further examined the concept of girl-child education to be all inclusive, some hindrances to effective girl-child education such as economic factors, sexual violence and abuse, political factors, the school environmental factors and socio-cultural and religious factors were highlighted. Included in the paper also was the implications that effective girl-child education would have on nation building such as poverty-reducing effects, improves health and nutrition, reduces inequality, reduces women's fertility rates, lowers infant and mortality rates and increases women's labour force participation rates and earnings. The paper finally recommended among others that government at all levels should give more attention to girl-child education, well to do individuals can contribute to girl-child education by giving them scholarship to study in higher institution, provision of school facilities and equipment that can ease their learning effectively as it contribute to nation building and the need to create more awareness for parents on sexual violence and abuse and this can be through radio and newspaper jingles and advertisement as well as periodical seminars and conferences.

Keywords: girl-child, education, implication, nation, building, 21st century Nigeria.

I. Introduction

In Africa, women are considered as men's properties or pleasure objects. They are also considered as a 'machine' meant for producing children. These situations have resulted in unfair treatment of women especially with regards to education of the male-child than the female child. In the traditional Nigerian society, there exist the believe that women are second class citizen (Enejere, 1991). The author further averts that gender inequality in Nigeria is promoted by religious and communal customs. Young girls particularly in Northern Nigeria are denied the benefit of education. This has given consequences for both the individual and the society at large.

Obinaju (2014) sees education as inalienable right of all irrespective of the person's circumstance. Education in its general sense is a form of learning which the knowledge, skills, values, benefits and habits of a group of people are transferred from one generation to the next through storytelling, discussion, teaching, training or research. Education has been described as the most important aspect of human development, a key to a successful living, especially girl-child education (Micheal, 2011).

Girl-child education is a catch-all term for a complex set of issues and debates surrounding (primary education, secondary, and tertiary and health education in particular) for girl and women. Denying the girl-child access to education implies making her a dysfunctional member of the society. Statistics show that many girls are not enrolled in school. The global figure for out of school children is estimated at 121 million, 65 million are girls with over 80 percent of these girls in sub-sahara Africa including Nigeria (UNICEF, 2007).

The concern of this paper is that despite the campaign by the Federal Government, United Nation Children Education Fund (UNICEF) and stakeholder in education to improve girl-child education in Nigeria, the level of discrimination on girl-child education is still high. Therefore, this paper seeks to redress the challenges that negate the effectiveness of girl-child education in Nigeria.

II. The Concept of Girl-Child Education

Within the context of education, many scholars have defined girl-child education in various ways. The National Child Welfare Policy (1989) as cited by Ada (2007) defines girl-child as a person below 14 years of age. Offorma (2009) defines girl-child as a biological female offspring from birth to eighteen (18) years of age. This period is made up of infancy, childhood, early and late adolescence stage of development. The girl-child is seen as a young female person, who would eventually grow into women and marry. She is conditional to look after the young ones the home and kitchen.

Girl-child education is a catch-all term for a complexity of issues and debates surrounding education (primary education, secondary education, tertiary education and health education for females (Okernmor, Ndit and Filshak, 2012). Girl-child education also includes areas of gender equality, access to education and it's connection to the alleviation of poverty, good governance, which are major ingredients

Author α: Department of Educational Psychology, Kogi State College of Education, Ankpa, Kogi State Nigeria.
e-mail: andrewomede@gmail.com
Author σ: Home Economics Department, Kogi State College of Education, Ankpa, Kogi State Nigeria. e-mail: agahiugrace@yahoo.com

© 2016 Global Journals Inc. (US)

in averting crime against women. Today's girl-child education is for her tomorrow's living. Afebendeugne in Ugwu (2001) defines women education as the education that would make a woman become aware of herself and her capacity to exploit her environment, and involves training in literacy and vocational skills to enable her become functional in the society. When maternal care is adequately provided for the girl-child the aims and objectives of education will be achieved.

However, current efforts including national and global programmes have been to target increased enrollment of the girl-child into the different levels of education in Nigeria. The federal government introduced the Universal Basic Education Programme to provide cheap and affordable education to all and sundry. Most if not all the state governments in Nigeria have also introduced free and compulsory primary and secondary schools for both male and female children in various states. Again most state governments have also passed the child rights and protection acts that will eliminate (or at least reduce) the withdrawal of the girl-child from school and to prevent parents or guardians from using their school age children to hawk or do endless labour activities. This is so important because it promotes girl-child education which chances nation building.

III. Girl-Child Education and Nation Building

Education is one of the most critical areas of empowerment for Women Education is central to development and improvement of the nation's welfare. It is a powerful 'equalizer', opening doors to all to lift themselves out of poverty. Below are the roles of girl-child education to Nation Building:

- Poverty Reducing Effects: Girl-child education can vitally contribute to the attainment of the Millennium Development Goals. While two of the goals pertain directly to education, education also helps to reduce poverty, promote gender equality, lower child mortality rate, protect against HIV/AIDS, reduce fertility rates and enhance environmental awareness (Mordi, 2008).
- Improve Health and Nutrition: According to Kiki (2010) education greatly benefits personal health particularly for girl-child, it profoundly affects reproductive health immunization rates. Education may be the single most effective preventive weapon against HIV/AIDS. If the issue of HIV/AIDS is rampart in a particular country, the force and economic growth will be affected. Again through the awareness of girl-child education, the rate of HIV/AIDS will be reduced to the barest minimum and this will have positive impact on Nation Building.
- Reduces Inequality: Education reduces illiteracy that is one of the strongest predictors of poverty. Primary education plays catalytic roles for those most likely to be poor, including girls, ethnic minorities, orphans, disabled people and rural families. By enabling larger members to share in the growth process, education can be the powerful tide that lifts all books (Okeke, Nzewi and Njoku, 2008).
- Reduces Women's Fertility Rates: Women with formal education are much more likely to use reliable family planning methods, delay marriage and child bearing, and have fewer and healthier babies than women with no formal education. This development enhances Nation Building.
- Lower Infant and Child Mortality Rates: According to Ocho (2005), women with some formal education are more likely to seek medical care, ensure their children are immunized, be better informed about their children's nutritional requirements, and adopt improved sanitation practices. As a result, their infants and children have higher survived rates and tend to be healthier and better nourished. If children survive through adequate medical facilities provided by a country, that aspect greatly enhance Nation building.
- Increase Women's Labour Force Participation Rates and Earnings: Girl-child education has been proven to increase income for wage earners and increase productivity for employers, yielding benefits for the community and society.

Despite these contributions to Nation building, women education has always been marginalized due to cultural practices and political decision. The effect of this situation has led to certain deprivations that have hindered women and girl-child in particular from maximizing their capacities in the development process of their communities. In Nigeria, as in other African countries, women are not held in high-esteem, consequently female education is seen as a wasteful venture as people think that the role of women is for procreation and confinement to the kitchen. This is the challenge of girl-child education today.

IV. Hindrances of Girl-Child Education

a) The following are some of the hindrances

- *Economic Factors:* Nigeria as an independent entity is undoubtedly characterized by very harsh economic conditions. This has resulted into scarce resources. As a result of this, choice has to be made between whom to send to school. Most often, it is the girl-child that remains at home. Due to poverty, girls get withdrawn from schools so as to help to supplement family income through hawking, trading or even working on the farm so as to support the family. In some cases, the girls are given out as house helps or even sent into early

© 2016 Global Journals Inc. (US)

marriage because of a huge bride price (UNICEF, 2007).

- *Sexual Violence and Abuse:* This also hampers the girls from going to school due to this fear of sexual violence, most parents deny their girl-children access to school.
- *Political Factors:* Despite the fact that Nigeria is signatory to various international conventions on the right of children, so far very little has been achieved in respecting children rights. The situation remains pathetic and serious. For instance, at the formation of the United Nations which is almost six (6) decades old, the precarious situation of the children worldwide became so obvious that it became necessary to establish UNICEF with special focus on the needs of the children around the world.
- *The School Environmental Factors:* Often most parents are scared of sending their female children to school in distance places and would rather keep them at home. According to Obinaju (2014), curricular, textbooks and other materials are usually gender-biased. She opines that right from childhood, girls are channeled into stereotyped traditional carrier in form of textbooks illustrations and stories consequently leading to the development of poor self image at a tender age. Also sexual harassment during educational pursuit create serious emotional and psychological strain on the girl-child.
- *Socio-Cultural and Religious Factors:* In most African societies, especially in Nigeria, the role of the girl as a wife and mother is conceived as the utmost priority not only by her parents, but also by the girl-child herself. However, in the Nigeria context, gender discrepancy in education is sustained by cultural factors. The wrong notion that her place is in the kitchen, to be seen and not to be heard have had very serious implications on the girl-child's ability at self-actualization. Obinaju (2014), notes that out of the 130 million children in LCDs without access to education, 81 million are girls. Also certain cultural and traditional practices like female circumcision, early marriages etc are to say the least unprogressive because they lead not only to absenteeism distraction, but also to eventual dropout of girls. Moreso, the ethnic and values of some religions do not help matters, as they are often perceived with tremendous suspicions.

V. Conclusion

The importance of education to a child and also for overall development of a nation cannot be over emphasized particularly the girl-child. Therefore, education is the right of every girl-child, a key to transforming her life and making her a responsible member of society. Without education, girls are denied the opportunity to develop their full potentials and play productive role in Nation Building. Although some efforts have been made to improve girl-child education in Nigeria, much still needs to be done if women must realize their potential and fully contribute to the political, socio-economic and technological transformation of the country.

VI. Recommendations

In order to overcome the challenges of girl-child education in Nigeria, the following recommendations are put forward:

- Government at all levels should give more attention to girl-child education. This is because if they are well educated, they will have chances of contributing to nation building.
- Well to do individuals can contribute to girl-child education by giving them scholarship to study in higher institutions, provision of school facilities and equipment that can ease their learning effectively as it contributes to nation building.
- The need to create more awareness for parent on sexual violence and abuse is imperative. This can be through radio and newspaper jingles and advertisement; as well as periodical seminars and conferences. More so, government should enforce law as regard to sexual violence and abuse in order to deter others from terrorizing the girl-child while in the school as it contribute to nation building.
- Nigeria negative attitude towards girl-child education should be discarded completely.
- UNICEF policies on equal rights to education should be strictly followed by government at all levels in Nigeria for the enhancement of Nation building.
- Government and non-government agencies should establish more boarding schools for girl-children to discourage parents from the notion of geographical distance, environmental hazards vis-à-vis the vulnerability of the girl-children.
- Nigeria society should not use socio-cultural and religious factors as yardsticks to relegate the girl-child to the kitchen. The education of the girl-child should be as important as that of the boy-child if not more important as peoples' opinion assert that when a woman is educated, a nation is educated.

References Références Referencias

1. Ada, N.A (2007). Gender power and politics in Nigeria. Markurdi: Aboki Publishers, P. 314.
2. Enejere, .E. (1991). Women and Political Education in D. O. Chizea and J. Njoku (eds). Nigerian Women and the Challenges of overtime, Lagos: Malthouse Press Ltd, P. 44-51.

© 2016 Global Journals Inc. (US)

3. Kiki (2010). Girl-child Education in Nigeria: Why not? Accessed on 1/11/2010 at http://www.amplifv-vourvoice.orglu/kiki/2010/3/10/girl/education in Nigeria why-not.
4. Michael, I. (2011). Emir Harps on Girl-child Education. Nigerian Tribune, No. 15, 107 5th January, P. 31.
5. Mordi, (2008). Founding Education in Nigeria: Surviving on a shoestring Budget. TELL Magazine, No. 44, November 3, p. 32-34.
6. Obinaju, Q.I. (2014). Gender Issues in teaching as profession, in E.N. Okpara (EN) Gender Issues in Education and Development. A Book of Reading (Vol. 8, pp 204-212) Enugu. University Trust Publishers.
7. Ocho, L.O. (2005). Issues and Concerns in Education and Life. Enugu: Institute of Development Studies, University of Nigeria, Nsukka P.24.
8. Offorma, G.C. (2009). Girl-child Education in Africa in Africa. Keynote Address presented at the Conference of University WOMWNE of Africa Held in Lagos, Nigeria, 16th – 19th July.
9. Okeke, E.A., Nzewi, U.M. and Ujoku, z (2008). Tracking School Age Children's Education Status in UNICEF A-Field States. Enugu UNICEF, P. 31.
10. Okenmor, G.A, Ndit, J.N. and Filshok, M.A (2012). The role of women education in conflict resolution and peace building in the present political dispensation in Nigeria Journal of Women in Colleges of Education 16 (2) p 71-77.
11. Ugwu, A.B.C (2001). Education for Special Target Groups Emene, Enugu: Fred-Ogah Publishers.
12. United Nations Children Education Fund (UNICEF) Factors Sheet (2007). Accessed on 1/11/2010 at http.//www.unicef.org/wcaro/wcaro Nigeria Fact Sheets Girl Education pdf.

© 2016 Global Journals Inc. (US)

Global Journal of human-social science: G
Linguistics & Education
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-460X & Print ISSN: 0975-587X

Trends and Prospects of Instructional Material Development and Delivery at the University of Zambia

By Mr. Kasonde Mundende, Mr. Francis Simui, Mr. Albert K Chishiba, Mr. Godfrey Mwewa & Boniface Namangala
University of Zambia (UNZA), Zambia

Abstract- The major concern addressed in this study is why the University of Zambia (UNZA), despite being the first national University with so much vast experience in excellence and service as its motto is and being the first institution to offer distance education (DE) in Zambia, has not commensurately grown to competitive and comparative massive enrollment numbers, increased number of programmes and courses and development and delivery of instructional materials in Open and Distance Learning (ODL). The need to introduce DE at UNZA, amongst many other compelling reasons could be attributed to the Lockwood Commission, whose report in 1963/64 recommended that degree programmes at the Institution should be available by distance study to suitably qualified candidates who might not be in position to attend the University education on full - time basis. This study is anchored on questions such as: (i) How does UNZA run distance education? (ii) How does UNZA develop instructional materials? (iii) How does UNZA deliver instructional materials to its distance students? (iv) What successes has UNZA scored in distance education mode of study? (v) Why is distance education at UNZA not expanding as expected?

GJHSS-G Classification : FOR Code: 130399

TRENDSANDPROSPECTSOFINSTRUCTIONALMATERIALDEVELOPMENTANDDELIVERYATTHEUNIVERSITYOFZAMBIA

Strictly as per the compliance and regulations of:

© 2016. Mr. Kasonde Mundende, Mr. Francis Simui, Mr. Albert K Chishiba, Mr. Godfrey Mwewa & Boniface Namangala. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http:// creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Trends and Prospects of Instructional Material Development and Delivery at the University of Zambia

Mr. Kasonde Mundende [α], Mr. Francis Simui [σ], Mr. Albert K Chishiba [ρ], Mr. Godfrey Mwewa [ω] & Boniface Namangala [¥]

Abstract- The major concern addressed in this study is why the University of Zambia (UNZA), despite being the first national University with so much vast experience in excellence and service as its motto is and being the first institution to offer distance education (DE) in Zambia, has not commensurately grown to competitive and comparative massive enrollment numbers, increased number of programmes and courses and development and delivery of instructional materials in Open and Distance Learning (ODL). The need to introduce DE at UNZA, amongst many other compelling reasons could be attributed to the Lockwood Commission, whose report in 1963/64 recommended that degree programmes at the Institution should be available by distance study to suitably qualified candidates who might not be in position to attend the University education on full - time basis. This study is anchored on questions such as: (i) How does UNZA run distance education? (ii) How does UNZA develop instructional materials? (iii) How does UNZA deliver instructional materials to its distance students? (iv) What successes has UNZA scored in distance education mode of study? (v) Why is distance education at UNZA not expanding as expected? (vi) How best could UNZA deliver its learner support services to its distance students? The study employed descriptive design and document analysis review. It is largely qualitative in nature. The major findings of the study show that UNZA being a dual mode institution depends on lecturers from various teaching schools to service distance education students; Institute of Distance Education (IDE), the Unit of the University with sole responsibilities of administration, organisation and coordination of DE programmes depends on lecturers employed under regular mode of study to develop instructional materials, conduct face-to-face contact sessions during residential schools, prepare assessment and examinations and mark them. IDE experiences some degree of resistance from mainstream faculty members in executing and implementing its distance education mandate. Furthermore, UNZA distance education over-relies on print media. At times, students in remote rural areas do not receive their study materials in good time, thus, compromising the quality of learner support services to them. The other finding is that management of student records is mostly done manually. Furthermore, currently, there is only one functional industrial machine for mass production of study materials which from time to time breaks down due to large volumes of materials to be produced. The other challenge is very low staffing levels for IDE. Currently, there are five academic members of staff and 33 supporting staff against slightly above 8000 students. On a serious note, UNZA does not have ODL policy, though in advanced stage of having one. Despite the aforementioned challenges, UNZA has made greater strides in maintaining its leadership in DE. For instance, the Institution has for the past two years, students completing their education through distance learning (DL) mode which was not the case previously as they had to transfer to full-time for their last two years of study. UNZA is also the leading institution in providing training, consultancy and advisory services in education to various institutions of learning in Zambia as regards ODL (including the use of ICT in ODL). Additionally, UNZA, though on a slow pace, is using e - learning platforms and its distance students access various information through the same platforms. More so, in addition to the Art-based programmes, UNZA is offering science-based programmes from the Schools of Medicine and Veterinary Medicine. Furthermore, UNZA, in collaboration with Zimbabwe Open University (ZoU), in 2014 introduced post graduate programmes at Masters and DPhil level in various fields of study. There are a lot of opportunities for UNZA to grow in terms of numbers and programmes as well as in providing quality distance education to its distance students. The Institution also has ICT infrastructure which is being supported by Zambia Research and Education Network (ZAMREN), and Centre for Information and Communication Technology (CICT) whose offices are housed at UNZA. In order for UNZA to realise its full potential in terms of capacity building, increase in programmes and courses, the researchers recommend the following: (i) UNZA-IDE should develop ODL Policy which will guide operations of DE, (ii) IDE should become more autonomous and employ lecturers solely to run DE and not depending on those employed for conventional/regular students, (iii) UNZA should decentralise its distance education operations so that registration processes and residential sessions/schools could

Author α: A/HEAD, Programmes Development and Production, Lecturer and Researcher, Geography Education Specialist, University of Zambia (UNZA), Institute of Distance Education (IDE), Lusaka, Zambia.
e-mails: kmundende05@yahoo.com, kasonde.mundende@unza.zm
Author σ: A/HEAD, Quality Assurance and Research, Lecturer and Researcher, Inclusive Education Specialist University of Zambia (UNZA), Institute of Distance Education (IDE), Lusaka, Zambia.
e-mail: francis.simui@unza.zm
Author ρ: Mathematics and Sciences Coordinator, Lecturer and Researcher, Statistics and Operational Research Specialist, University of Zambia (UNZA), Institute of Distance Education (IDE), Lusaka, Zambia. e-mail: achishib@unza.zm
Author ω: HEAD, Learner Support Services, Lecturer and Researcher, Instructional Technology Specialist, University of Zambia (UNZA), Institute of Distance Education (IDE), Lusaka, Zambia.
e-mail: Godfrey.mwewa@unza.zm
Author ¥: Professor, Director, Lecturer and Researcher, Microbiology/Immunology Specialist, University of Zambia (UNZA), Institute of Distance Education (IDE), Lusaka, Zambia.
e-mail: b.namangala@unza.zm

© 2016 Global Journals Inc. (US)

be done in provincial centers, (iv) UNZA - IDE should strengthen ICT infrastructure to allow the provision of providing modern methods of delivery such as moodle, and (v) UNZA should have a separate DE calendar from the regular students' academic calendar.

I. Introduction

Distance Education (DE) is another mode of study. Countries world - wide have seen the benefit of using this mode of study to firstly increase the enrolment levels of those pursuing education. Secondly, it is a mode of study which provides education to qualified people who are already in either formal or informal employments and may not have time to attend a full time programme. Thirdly, it is a suitable mode of study for busy people who can still upgrade themselves, whilst working. It is referred to as flexible mode of study because it is pro – learner. With time, it gained another dimension of Open and Distance Learning (ODL). Most ODL systems have a philosophy that aims to: remove barriers, and allow learners to study what they want, and when they want and where they want (Commonwealth of Learning (COL), 2005). ODL systems use technology to mediate learning such as printed workbooks, audio cassettes, radio and the web. By implication, there is no one method for providing ODL (COL, 2005).

Some American higher education administrators and state politicians recommended distance learning (DL) as one which would increase solve problems related to maintaining access to education, meeting increased demands to education and efficient use of technology amongst many other aspects (Basom and Sherritt, 1994).

a) Background to the study

The University of Zambia (UNZA) is the first state University in Zambia. It may be referred to as the 'mother' of all public and private Universities in the Country. It is also the first University to offer distance education to suitably qualified candidates since its inception in 1967. From inception, the institution has passed through various developmental stages. Right from the inception, the institution started running DE partly through the recommendation by the Lockwood Commission, whose report in 1963/64 recommended that degree programmes at UNZA should be available by distance study to suitably qualified candidates who might not be in position to attend the University education on full- time basis. In line with the Commission's recommendation, an autonomous Department of Correspondence Studies (DCS) was established in 1966. DCS was meant to coordinate distance learning (DL) programmes which were launched in March 1967, with about 200 registered students. In 1975, DCS became part of the Centre for Continuing Education (CCE), which later in 1994, became the Directorate of Distance Education (DDE). With the expansion programmes on board, fifteen (15) years later in September 2009, DDE changed its name to Institute of Distance Education (IDE) and currently has about 8000 registered students. For now, the organisation, administration and coordination of DE courses at UNZA is the responsibility of IDE, but all tuition is delivered by members of academic staff of various Schools of the University. Members of staff of the Schools offering courses by distance teaching prepare all study materials and assignments in accordance with approved course outlines. From inception till about 2012, students were spending two years under distance mode and completed their other two years under full – time mode of study. This, to some extent, made some students drop out because their employers did not allow them to obtain leave for two years. Further, students were only taking four (4) courses and that made them to take longer than their counter parts in full – time to complete their study programmes. This is no longer the practice as of now because students are now completing their studies under distance mode and are taking at least 8 courses per academic year, just like their full – time counter parts. This development has been made possible because most third and fourth year study materials (modules) are now available in almost all courses, which was not the case before 2012. UNZA maintains excellence and service for its students both full – time and distance in that there is no significant difference between the quality of the two processes (Basaza, Milman and Wright, 2010). UNZA is a mixed institution with a total number of registered students of about 26000. With 48 years of existence and experience, one wonders why UNZA, has not been a leader in student enrollment numbers, programmes and courses in comparison with some recently established institutions locally and within the region.

b) Problem Statement

Although UNZA started offering distance programmes some 48 years ago, on comparison basis, with newly established institutions running distance programmes locally (Such as Kwame Nkrumah University, Kabwe) and within the region (Such as Zimbabwe Open University, Zimbabwe), UNZA has not commensurately grown in terms of enrollment levels, increased number of programmes and courses as well as development and delivery of instructional materials. Kwame Nkrumah University commenced its distance education programmes in 2012, and by today it has grown to about 4 000 DE students. ZoU which started in 1995, has about 20, 000 DE students, whilst UNZA with 48 years of existence has slightly above 8000 students. The question under probe is despite 48 years of distance education existence, why is UNZA not greatly expanded in terms of number of student enrollment,

© 2016 Global Journals Inc. (US)

programmes and courses as well as development and delivery of instruction materials?

c) Aim of the Study

The study aims at bringing out salient trends and prospects in the running of DE at UNZA. It also aims at how to turn challenges into opportunities so that UNZA can have quality massive enrollment levels as expected in ODL.

d) Research Questions of the Study

This study addresses the following questions:

- How does UNZA run distance education?
- How does UNZA develop instructional materials?
- How does UNZA deliver instructional materials to its distance students?
- What successes has UNZA scored in distance education mode of study?
- Why is distance education at UNZA not expanding as expected?
- How best could UNZA deliver its learner support services to its distance students?

II. Literature Review

There has not been a universally accepted definition of distance education, but a number of authorities have endeavoured to define the concept in various ways. Homberg (1995: 2) defines DE as being 'characterised by non – contiguous communication between the supporting organisation and its students'. Homberg further highlights that there are two modes of communication namely: one – way traffic and two - way traffic. In one - way traffic, which is also referred to as stimulated communication, pre – produced course materials are sent from the supporting organisation and involves students in interaction with texts, recordings and data bases. In two - way traffic, which is also called real communication, this is done between students and their supporting organisation, in writing, on the telephone, and by fax or mail. Keegan (1990:44) defines DE as 'the use of technical media – print, audio, video or computer to unite teacher and learner and carry the content of the course'. Keegan (ibid) further defines distance education as 'the quasi – permanent separation of teacher and learner throughout the length of the learning process (this distinguishes it from conventional face to face education). 'By implication, DE has to do with offering education to qualified human beings, regardless of who they are, where they are, what they do, but are willing to learn, upgrade and still continue with their other responsibilities with minimum resistance and disturbance'.

DL world-wide offers promises, problems and possibilities. Horgan (1998:1) observes that "many Universities are feeling the pressure to control their costs, improve quality of instruction, focus on customer needs, and respond to the competitive pressures". DL technologies have the potential to assist in solving these problems. Basom and Sherritt (1994:4) surveyed higher education administrators and state politicians to find out what they thought would be the major problems facing American higher education in the next millennium. The answers they most often received were: "meeting increased demands at a time of decreased resources, increasing or maintaining access, using technology more efficiently, and sharing resources across state lines so that colleges won't have to be all things to all people". From their conclusion, DL seems to address all of these issues. Dibiase (2000) subscribes to Basom and Sherritt (1996) that Administrators hope that distance learning methods will help make higher education more cost-effective. Valentine (2008) outlines quality of instruction, cost effectiveness, misuse of technology, the role of the technicians, problems with equipment, attitudes towards DL as problems of DL. He further highlights promises of DL which include (i) class size increases while the overhead remains the same, (ii) through the use e-learning techniques, ability to reach those that would by any other means be unreachable. Countries like China use a radio and television delivery system to serve 1.5 million students, two thirds of which are in a degree program.

As regards the instructional design process in DL, there is no teacher because he/she is replaced by a combination of learning materials and expensive tutors. Due to the fact that tutors meet their students for a short period of time, learning materials themselves would among many other important aspects define what is to be learnt, provide information and give examples, equivalent to what a conventional teacher does in class (COL, 2005).

From the African education perspective, the long term aims of the Ministry of African Education have been to increase the number of Africans qualified to play a full part in administrative and social services, in industry and commerce and in public life generally (Kelly, 1999). The factors that led to the development of DE at the time of Zambia's independence included response to the country's need for human resource development at the time of political independence and the demand for this level of education from people who had got into employment without university education. DE was, therefore, identified as a means of expanding enrolment for university education (Siaciwena, 1988).

According to the University of Zambia's Strategic Plan (2001: 24), the distance education program was the means of

> Providing wider access to arrange of entrants through diversified
>
> and flexible formats of study, to cater for those unable to participate

© 2016 Global Journals Inc. (US)

in regular schemes of study, and through the effective utilization

of the existing ICT infrastructure especially its use in DE

Being flexible and open to all (UNESCO.-org/edu/en; Commonwealth of Learning and Asian Development Bank, 1999), Zambia increased the number of its citizens who entered school at that time.

III. Research Methodology

The researchers employed the descriptive survey design. In educational research it is argued that descriptive survey is a method of research that describes what is seen over and beyond (Babbie, 1997). This method allowed researchers to get the rightful information according to what they have seen happening at UNZA as regards distance education operations since its inception. Descriptive studies result into fact finding, formulation of important principles of knowledge and solution to significant problems. Additionally, they involve measurement, classification, analysis, comparison and interpretation of data (Kombo and Tromp (2006). The study was largely qualitative and document analysis review was employed, using secondary data. It is believed that qualitative methods can be used to bring out what may be 'hidden' and what is little known (Strauss and Corbin, 1990). The researchers reviewed various literature on the subject at hand. Data were collected from literature on UNZA activities as regards distance education operations. As for data analysis researchers used graphs as well as themes to interpret the information.

IV. Research Findings

The study brings out various results which comprise challenges and successes UNZA has experienced since its inception. Firstly, though UNZA started offering distance programmes some 48 years ago, on comparison basis, with newly established institutions running distance programmes in Zambia and within the region, the institution has been facing serious challenges as regards learner support systems largely in the development and delivery of the instructional materials.

a) *UNZA challenges in running Distance Education Mode*

According to Siaciwena (2000b), UNZA for some time has not been able to reach out to as many people as it would have potentially done due to, firstly, the capacity to handle an increasing numbers of students which exerts lots of pressure on material development and the need to provide an effective learner support service; secondly, the inadequate means of communication with students which has been largely on mobile phones to those who are accessible because some students are in rural Zambia where accessibility is difficult. Other forms of communication are through letters which are posted through Zambia Postal Services, which at times are slow and letters or study materials take long to reach and at times get lost along the way.

Secondly, being a dual mode institution, UNZA - IDE Unit does not have its own lecturers. The University of Zambia uses what Peters (2002) calls the mixed – mode of distance education. In this approach, courses are developed on the same principles as those for conventional face - to - face students so that distance students follow the same curriculum and are subjected to the same performance requirements as the conventional students. The Institute depends on the same lecturers from various teaching schools of the University, who are already overwhelmed with teaching loads from Regular, Parallel and Fast Track students. These are the same lecturers who develop instructional materials; provide guidance, conduct face-to-face contact sessions during residential schools; prepare assessment and examinations and mark them. This arrangement where same staff teaches both conventional and distance students could be strenuous on the part of the staff and it is likely that the staff may not devote equal time to distance teaching (Siaciwena, 2000b). At UNZA, as a matter of procedure, a program will not be offered under the distance learning mode, until it has been offered to conventional students for at least one year (Siaciwena, 1988; Chifwepa, 2006).

Thirdly, there is also a degree of resistance from mainstream faculty members. Some of them treat distance students as inferior and secondary to conventional students. This creates some 'stigma' to the prospective distance students who feel their degree qualification would be inferior to the ones obtained by their face – to – face counterparts.

Another finding is that UNZA - IDE over-relies on print - media at the expense of e - media. This perhaps is far less technological in comparison to what countries like China have ventured into, such as using a radio and television delivery system to serve 1.5 million students, two thirds of which are in a degree program (Valentine, 2008). Over relying on print – media has its own disadvantages in that, at times, students in rural Zambia do not receive their study materials in good time, mostly due to poor postal services challenges, thus compromising on the quality of learner support services on them. One major challenge facing UNZA is lack of an ODL policy for all these years, making it difficult to execute certain operations effectively.

The other problem has to do with inadequate academic and administrative staffing levels. As far back as the year 2000, UNZA has been having a skeleton staffing to administer distance education mode of study (Siaciwena, 2000b). Currently, the IDE has Thirty – Eight (38) permanent members, *five* (5) academic members of

© 2016 Global Journals Inc. (US)

staff and *thirty - three* (33) support staff members against Eight (8000) registered students. This disparity causes much pressure on these few members in all spheres.

One more challenge has to do with management of student records. Largely, records and materials for distance students at UNZA have been maintained manually in the IDE (Siaciwena, 2000b, Chifwepa, 2006). The picture as observed by Siaciwena and Chifwepa some fifteen and nine years ago respectively, has not changed much for UNZA – IDE. Though there is a slight improvement through CICT, the general student record system of the University is inadequate in providing the type of the information that the IDE needs for the management of the distance education programmes.

Last but not the least, UNZA – IDE depends on one industrial machine to mass produce not less than seventy thousand (70 000) study materials for students, besides other productions made for the Unit. This one machine is ever overwhelmed and breaks down form time to time. This creates a shortage of study materials which are supposed to be issued out to students during their registration at the beginning of the academic year. Compounded by this, study materials are centrally distributed. As observed by Siaciwena some fifteen years ago (Siaciwena, 2000b), UNZA – IDE, still distributes its study materials to its students from the Great East Road Campus, Lusaka. With this 'gloomy' picture, there is need for UNZA to procure some more industrial machines to meet the increasing demands of enrollment levels.

b) *UNZA Successes in Running Distance Education Mode*

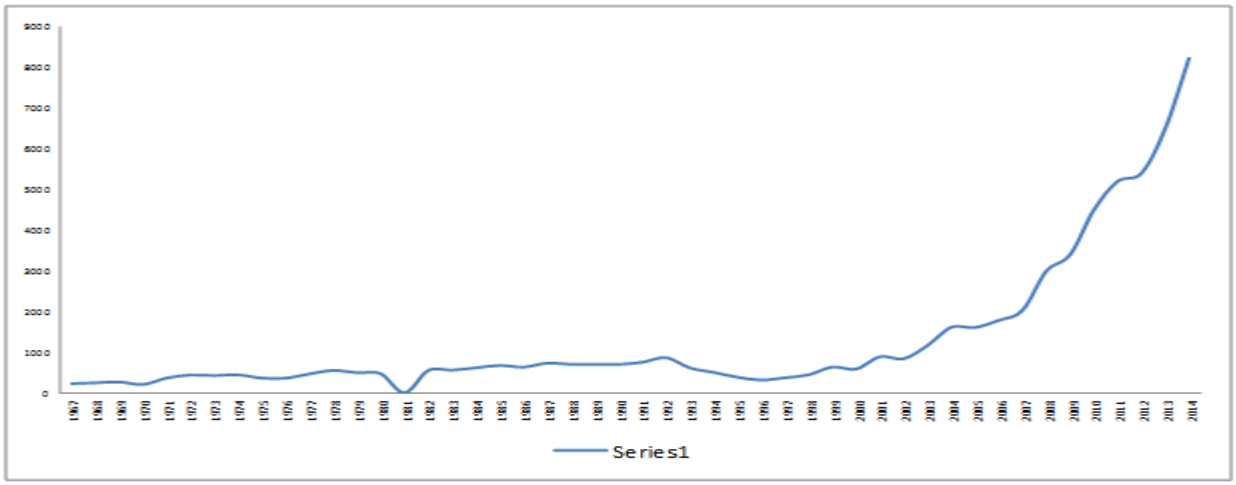
Despite the aforementioned challenges, UNZA has made greater strides in maintaining its leadership in DE. For instance, UNZA has for the past two years students who complete their education through distance learning mode which was not the case previously, as they had to transfer to full-time for their last two years of study.

Secondly, UNZA is also the leading institution in providing training, consultancy and advisory services in education to various institutions of learning in Zambia as regards ODL (including the use of ICT in ODL). This far, UNZA – IDE has facilitated in capacity building a total number of *twenty - four (24) institutions* across Zambia as well as in the management of ODL. The other area has been in equipping those who have been capacitated in Instructional Material Designing, Development and Production. This is besides an in – house training of UNZA lecturers in instructional designing and development, whose number currently reaches over *five hundred (500) lecturers*.

Thirdly, the University, though on a slow pace, is using e - learning platforms and its distance students access various information through this platform. To support e – learning, UNZA has ICT infrastructures which are being supported by Zambia Research and Education Network (ZAMREN) as well as Centre for Information and Communications Technology (CICT), whose offices are housed at UNZA.

The other finding is that in addition to Art-based programmes, UNZA is running science-based programmes by the Schools of Medicine and Veterinary Medicine. Additionally, in 2014, UNZA, in collaboration with Zimbabwe Open University (ZoU), introduced Post graduate programmes (that is, Masters and doctoral programmes) in various fields of study.

The other development is, UNZA has increased in enrolment levels (Chifwepa, 2006). Fig 1. Shows how UNZA – IDE has grown in terms of numbers since its inception, 1967.



(IDE Statistics, 2015)

Figure 1 : Distance Education student enrolment trend analysis, UNZA

© 2016 Global Journals Inc. (US)

V. Conclusion

The study on the *'Trends and Prospects of Instructional Material Development and Delivery at the University of Zambia'* concluded that UNZA has not grown as expected in enrollment levels, programmes and courses, development and delivery of instructional materials to distance students despite being in existence for the past 48 years. There are factors to that effect which border on: negative attitude by some academic members of staff from teaching schools which service IDE; over reliance on print materials by IDE and minimal use of ICT platforms by DE students, very low staffing levels in IDE Unit; banking on one industrial machine to produce over 70 000 study materials as well as having no ODL Policy as some of the major challenges impeding the expected expansion. The study also concludes that there are however, greater opportunities for UNZA to grow in terms of numbers and programmes as well as in providing quality distance education services to its distance students. Such opportunities include, ICT infrastructure such as ZAMREN and CICT housed within UNZA; increased number of courses taken by distance students from 4 to 8 courses; the completion period of 4 years, equivalent to their counter parts in the conventional face – to – face mode of study among others. The other opportunity is UNZA still commands its leadership in service and excellence in the provision of ODL programmes in the entire country. Last but not the least, the University Management supports the IDE expansion programmes. This fact is embedded in the UNZA Strategic Plan (2013 – 2017:37) which states that 'UNZA will provide Open Distance Learning as an Option for Mass Teaching, Learning and Research'. Currently, IDE was granted the status of being an admitting school, just like any other School in the University. What is required is for UNZA to work on its DE challenges, which authors believe the institution can do with minimal impediments. UNZA still commands excellence and service in the country and can soar above its challenges and maintains its leadership in distance education provision. There should be a deliberate move and policy by UNZA to expand its numbers all round, through distance learning mode. The University can not afford to lag behind in DE being the 'mother' of all Universities in the Country, facing various competitors in DE mode as well as being the first one to commence distance education programmes.

VI. Recommendations

In order for UNZA to provide the much needed learner support services to its distance students, and increase in terms of the needed human capacity through distance learning mode as expected, we recommend the following:

- UNZA should have an ODL Policy which will guide operations of DE,
- IDE should become more autonomous and employ lecturers solely to run DE and not depending on institution's teaching schools lecturers who are already overwhelmed with conventional loads.
- UNZA should decentralise its distance education operations so that registration processes, distribution of study materials as well as residential sessions could be done in provincial centers,
- UNZA should strengthen Information Communication Technology (ICT) infrastructure to allow the provision of modern methods of delivery to its distance students, and
- UNZA should have a separate DE academic calendar from the regular students' academic calendar.

References Références Referencias

1. Babbie, E., (1997). *The Practice of Social Research.* New York, NY: Wadsworth.
2. Basom, M., and Sherritt, C., (1992). "Higher education problems in the twenty-first century: A survey of higher education administrators and politicians". Paper presented at the Annual Conference for International Higher Education Administrators, Nice, France.
3. Chifwepa, V., (2006). Development of a Model Plan for the Application of Information Communication Technologies in Distance Education at the University of Zambia. Thesis (Ph D), University of Zambia (Unpublished).
4. Commonwealth of Learning and Asian Development Bank (1999). *Designing Materials: For Open and Distance Learning*. Asian Development Bank: Manila.
5. Dibiase, D., (2000). "Is distance education a Faustian bargain"? *Journal of Geography in Higher Education*, 24(1), 130-136.
6. Homberg, B., (1995). *The Sphere of Distance Education Theory Revisited.* Zentraies Institute fur Fernstudienforschung: Hagen.
7. Horgan, B., (1998). "Transforming higher education using information technology: first steps". (On-Line). Available: http://microsoft.com/education/hed/vision.html
8. Keegan, D., (1990). *Foundations of Distance Education*. Routledge: London.
9. Kelly, M.J., (1999) *The Origins and Development of Education in Zambia – From Pre – Colonial Times to 1996*. Image Publishers Limited: Lusaka.
10. Kombo, K. D., and Tromp, A. L. D., (2006). *Proposal and Thesis Writing: An Introduction.* Paulines Publication Africa: Nairobi.

© 2016 Global Journals Inc. (US)

11. Lock wood, J., (1964). Report on the Development of a University in Northern Rhodesia. Lusaka: Government Printer.
12. Peters, O., (2002). Learning and Teaching in Distance Education [On line]. Available from: http://www.fernuni – hagen. De/ZIFF/V2 – ch40a.-htm. [Accessed on 11/08/2015]
13. Sherritt, C. (1996). "A fundamental problem with distance programs in higher education". (Opinion paper no. 120). Viewpoints. (ERIC Document Reproduction Service No. ED 389 906).
14. Siaciwena, R.M.C., (1988). 'A study of Distance Education at the University of Zambia with Special Reference to the Effectiveness of Degree Courses'. Thesis (Ph D), University of Wales. (Unpublished).
15. Siaciwena, R.M.C. (2000b). Management of Open and Distance Learning in Africa: The Case of the University of Zambia (Unpublished).
16. Strauss, A.L., and Corbin, J., (1990). Basics of Qualitative Research Procedures for Developing Grounded Theory. Newbury Park, CA: Sage.
17. The University of Zambia Strategic Plan 2001.
18. The University of Zambia Strategic Plan 2013 – 2017.
19. UNESCO.org/edu/en
20. Valentine, D., (2008). *Distance Learning: Promises, Problems, and Possibilities*. Oklahoma.

© 2016 Global Journals Inc. (US)

This page is intentionally left blank

© 2016 Global Journals Inc. (US)

Global Journal of human-social science: G
Linguistics & Education
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-460X & Print ISSN: 0975-587X

A Sociolinguistic Study of Choosing Names for Newborn Children in Jordan

By Abeer Harb Al-Qawasmi & Prof. Fawwaz Al-Abed Al-Haq

Yarmouk University, Jordan

Abstract- This study aims at the study of newborn names in Jordan of a sociolinguistic perspective. This study tries to detect the difference in naming newborns in Jordan over the decades - from the seventies to 2015 due to the result of some factors that may have affected the Jordanian society, whether historical, religious and/or social. The data necessary to complete the study was obtained from the Civil Status Department and the Department of Statistics. The data obtained consisted of names of both sexes during the time period from the seventies until the early year of 2015, a random sample of personal names within the same family were also provided. The data was analyzed quantitatively. The study revealed that there is a clear change in the choice of newborn names - male and female - in Jordan, whether a change in sounds or in morphemes. In specific, names during the seventies were strongly linked to the culture and the values, religious or social, in which the people believed in. During the eighties and nineties, names were associated with certain social values, however, some names were shown to be affected by urbanization or modernization.

Keywords: naming practice, newborns, culture, sociolinguistic.

GJHSS-G Classification : FOR Code: 200499

Strictly as per the compliance and regulations of:

© 2016. Abeer Harb Al-Qawasmi & Prof. Fawwaz Al-Abed Al-Haq. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http:// creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

A Sociolinguistic Study of Choosing Names for Newborn Children in Jordan

Abeer Harb Al-Qawasmi [α] & Prof. Fawwaz Al-Abed Al-Haq [σ]

Abstract- This study aims at the study of newborn names in Jordan of a sociolinguistic perspective. This study tries to detect the difference in naming newborns in Jordan over the decades - from the seventies to 2015 due to the result of some factors that may have affected the Jordanian society, whether historical, religious and/or social. The data necessary to complete the study was obtained from the Civil Status Department and the Department of Statistics. The data obtained consisted of names of both sexes during the time period from the seventies until the early year of 2015, a random sample of personal names within the same family were also provided. The data was analyzed quantitatively. The study revealed that there is a clear change in the choice of newborn names - male and female - in Jordan, whether a change in sounds or in morphemes. In specific, names during the seventies were strongly linked to the culture and the values, religious or social, in which the people believed in. During the eighties and nineties, names were associated with certain social values, however, some names were shown to be affected by urbanization or modernization. And with the beginning of 2000 up to 2015, peoples directions towards naming newborns changed due to the advent of globalization, associating with development and urbanization, and moreover, the influence of different cultures on the community.

Keywords: naming practice, newborns, culture, sociolinguistic.

I. Introduction

Naming is an essential and universal process, which people use to identify what surrounds them as a way of communication around the world. The word 'name' has been defined as "a word or a group of words that suggest a certain entity whether it's a real or fictitious" (Encyclopaedia Britannica, 1926). The definition of the word "name" in Oxford Advanced Learner's Dictionary is "A word or set of words by which a person or thing is known, addressed, or referred to." Everything around us has to be identified by giving it a suitable name, to enable us to distinguish between these things. Many scholars and linguists have focused on the investigation of how names can reveal cultural insights, linguistic features, and the historical background of a society. For example, Lyons (1977) states that "the functional semantic of words has been seen as that of naming, as far as we can sign or indicate the history of linguistic estimation". According to Bean (1980:306), states that there is a connection between a name and its bearer, which is created after choosing such a name. Also, Palmer (1981) shows that names are words used as a signifier and what objects referred to as a descriptor.

For human beings, a name is definitely the most important rights for human being that should carry an esthetical characteristic and have an acceptable meaning in society, so everyone should have a good name. The names which belong to humans are personal names, these are either person's name, family name or even person's nicknames. Mehrabian and Piercy (1993) mention that the name gives an impression, either positive or negative such as (Whʃa) وشحش , and has a corresponding impact in social interaction. Naming one's child with a pleasant name is one of the Islamic principles.

So, muslim parents are forced to choose a good name for their children. The Prophet Muhammad – peace and praise be upon him- said, "To whoever is born a boy and he names him Mohammad solely for the love of one and for the blessing name then both he and his son", because names affect the person's character (Raafat:2004). Because personal names give the confidence to their bearers, that tell everybody all over the world who you are, and may tell the occasion and the origins behind choosing such names.

Names are studied from a semantic point of view. Not all names have a meaning, when they have a meaning, they tend to have a positive impression (Zuraiq, 1999). As stated by Crystal, sound and meaning are hard to separate from each other (Crystal, 1995).

Socio-cultural background, attitudes, beliefs, and physical environment are non-linguistic factors, according to (Zuraiq, 1999). Every society has its own traditions when it comes to naming their newborns. When parents name their children, they tend to avoid names which may violate their social norms, and they seem to stick to certain names used within the same family, such as, naming their children after their grandparents, uncles, etc. Finally, as Rosen house (2002) states, naming is an act that reveals many factors, such as, traditions, hopes, feelings, fears, and daily events.

In Islam, the most beloved names to Allah are the following two: Abdullah and Abd-Ar-Rahman, as

Author α: Department of English Language and Literature, Faculty of Arts, Yarmouk University, Irbid – Jordan.
e-mail: abeerqawasmi26@yahoo.com
Author σ: Department of English Language and Literature, Faculty of Arts, Yarmouk University, Irbid – Jordan. e-mails: fawaz_m@ttu.edu.jo, fawaz_m@yu.edu.jo

© 2016 Global Journals Inc. (US)

stated in the Hadith narrated that Ibn Umar (RA) said: Allah's Messenger (PBUH) said, "The most beloved of names to Allah, the Mighty and Sublime, are: Abdullah and Abd-Ar-Rahman."

(Hadith No. 4949, Book of Etiquette, Sunan Abu Dawud, Vol. 5; Hadith No. 2833, Chapters on Manners, Jami' At-Tirmidhi, Vol. 5; Hadith No. 3728, Chapters on Etiquette, Sunan Ibn Majah, Vol. 5).

As Muslims, we must be careful enough to understand the process of naming a child. As a name for a child is sacred to him/her, so let it be good. When naming a child, the parents should take into consideration what the Glorious Qur'an and the Sunnah say.

The word "name" appears many times in The Glorious Qur'an. The first one when Allah taught Adam the names, which maybe the names of all creatures; humans, animals, trees, etc.

"ءلؤه ءامسيلب ينبئخأ لقف ةكئلملا ىلع مهضرع مث الهك ءامسلأا مدآ ملعو" (قرةبلا قروس : 31)"نقيدص متنك نإ

and –taught adam the-names all then Showed+them on the-angels and –said tell+me with names these if were you truthful (Al-Baqara, 2: 31)

"And He taught Adam all the names, saying: Inform me of the names of these if you are truthful." (Al-Hayik1995).

The second time, when Imran's wife speaks to Allah, after giving birth; "ىَهْنُأ ىِّنِإَو" (Aal-e-Imran, 3:36),

"I have –named her Mary" (Salahuddin, 1999).

In addition, when Allah brought to Zechariah a good news that he will have a boy whose name is John.

"ىيحي هسمٱ هل كرشبن لإ كئرشبن" (مريم روس : 7)"

"O Zakariyya (Zachariah) verily, We give you glad tidings of a son, whose name will be Yahyâ (John)". Al-Hilali and khan (2004, 1427) (Maryam, 19:7).

Akinnaso (1980) states that a personal name is a type of proper names. The process of personal naming have a main target, which is a system of symbols, that reveals the identity of the personal names.

The child has been given his/her name after birth; the name may be related to particular social factors. For example; Yassin (1978) states, "the names in Kuwait may indicate the social situations". He also adds that the names may refer to historical events, for example, sea objects, plants, and precious metals.

According to our Arab societies, the names of new babies are chosen by the parents or by close relatives, and those names can reveal the circumstances behind choosing such a name (Bean, 1980).

In the USA, the parents' political ideology can be seen through their children's names, Pappas (2013) states that parents political status, liberal or conservative, can be seen through their choice of baby names. She elaborates by saying that the sounds of liberal and conservative names vary. Parents in liberal neighborhoods are more likely to choose softer, more feminine sounds, such as "L," and soft-A endings for their babies' names, such as Liam, Ely, Leila, Ella and Sophia while conservative parents tend to pick names with more masculine-sounding K's, B's, D's and T's, such as Kurt.

Jayaraman (2005) states that names in India are chosen depending on the religion and region which the person is from. Elaborating that personal or first names such as Rasheed, Krishna, and David, and surnames such as Khan, Pradhan, and Das convey a person's religious. Therefore, it is possible to distinguish a Christian (David Das) from a Muslim (Rasheed Khan) from a Hindu (Krishna Pradhan). Furthermore, names in India also indicate whether the person comes from northern or southern India.

II. Methods and Procedures

The sample of this study consists of names of all registered newborn, male and female, children in Jordan during the period of the 70s to the early year of 2015. It also consists of names of newborn children within the same family, i.e., siblings, father/mother, grandfathers/mothers, uncles/aunts, etc. The selection of the families was random.

To fulfil the purpose of this study, the data which needed to conduct this study was collected by two methods. First, the researcher gathered data from the database at the Department of Statistics, listing the most frequently birth names by gender for the years 2009, 2010, 2011, 2012, and 2013.

Second, the researcher collected data from the Department of Civil Status, listing all personal names registered from the year 1980 to 2015. Furthermore, the researcher collected some additional data related to personal names within certain families.

The researcher applied the following methods in classifying and analysing the data related to the study of naming newborns in Jordan, in order to ease the understanding of how and why names are chosen.

- The researcher classified the names of newborn into three eras; from seventies, eighties to nineties, 2000 to 2015.
- The researcher analyzed how sound pattern of names has changed during time.
- The researcher analyzed names chosen within the same family.
- The researcher also showed the most frequent birth names, by gender, for the last 5 years after all the circumstances that have happened during these years.

© 2016 Global Journals Inc. (US)

III. Results and Discussion

For the purpose of the study, personal names were gathered from official documents provided by the database at the Department of Civil Status and from database at The Department of Statistics in Jordan.

Religion, naming after relatives, politics, naming according to surrounding environment, stressing certain times and places, and borrowing are the factors behind the sociolinguistic. So, studies of personal names have found that some personal names reveal at least one of these factors when the baby was born.

These names were categorized by the researcher into three main generation groups; The 70s, the period between the 80s and 90s, and the period between the year 2000 and the 2015. The researcher classified the names into these particular generations according to the similarity of the political, economical, social, and environmental events which have happened within each generation. Then, the identified names were compared in relation to the events, circumstances, and religion into which the children were born.

a) The Choice of Newborn Names during the 70s.

Under this heading, based on the analysis of the gathered data, the researcher discusses the data concerning male newborn names and female newborn names that were chosen for children during the 70s.

This period of time witnessed many historical, political, religious and social events, which had not only affected the Jordanian society, but also the choice of male and female newborn names.

i. *In Relation to Historical Events*

Due to the strong relations between the Jordanian and the Palestinian societies, many events such as the displacement and expulsion of Palestinians from their homeland and the following the Battle of Karameh have had a great effect on Jordanian society, which helped in the appearance of new names carrying the meanings of struggle and Jehad. In Arabic, the word jihād is a noun meaning "to strive, to apply oneself, to struggle, to persevere". Terminological definition: If called often go out to fight the unbelievers, stubborn, and warriors, and apostates, and prostitutes, and towards them, and its purpose is to Glorified the word of Allah.

During the 70s, the names which identified by the researcher, were mostly Jehadi names for both females and males equally. The most frequent names for males are /nidha:l/نضال (3.7%), /ǰiha:d/جهاد (2.9%), /na:ṣr/2.8) ناصر %), /fa:rs/فارس (2.6%), and for females /fida:ʔ/ فداء (3.9%), /kifa:ħ/ كفاح (6.4%), /ʔentiṣa:r/انتصار (5.9%), /naǰa:ħ/ نجاح (4.1%). The reason behind the choice of such names is considered to be the result of the events and circumstances which occurred before and during this period. Other historical events effecting the choice of newborn name in Jordan were the Six-Day War, known as النكسة, an-Naksah, when the Arab people tried to liberate Palestine, and The Yom Kippur War, حرب يوم خيبر, Ramadan War, or October War.

ii. *In Relation to Religion*

In relation to religion, Jordan is an Arab Islamic country, so we talk about an Islamic society which follows the Islamic precepts in every work of life. Names, as mentioned before, are considered a great thing in Islam. In Islamic sociolinguistics, parents can name their male newborn after Prophets and messengers. The most common male names are /muħamad/ محمد, /ʔħmd/ أحمد, /maħmwd/ محمود which are variant appellations of the Prophet Muhammad's name, peace be upon him. Among the female newborns, the following religious names are also popular, /fa:ṭma/فاطمة after the Prophet's daughter and /ʕa:ʔʃa/عائشة after the Prophet's wife, though not as frequent as the male names. These common names which, carry the Arabic-Islamic features, whether for male newborns or female newbornsare still chosen till this period of time. Such as the female names /fa:ṭma/فاطمة (12.5%), /muħamad/ محمد (21.6%) , /ʕa:ʔʃa/ عائشة (3.5%), /ʔħmd/(%11.0) أحمد. Table 1: shows the most frequents names chosen for this period of time.

Table 1 : Names and frequency during the 70s

Female names	Name Frequency	Rate%	Male names	Name Frequency	Rate%
كفاح kifa:ħ	724	6.4	نضال Nidha:l	796	3.7
فداء fida?	450	3.9	جهاد ǰihad	712	2.9
انتصار ʔenteṣar	683	5.9	ناصر na:ṣr	604	2.8
نجاح naǰa:ħ	468	4.1	فارس fa:rs	563	2.6
عائشة a:ʔʃaʕ	401	3.5	محمد muħamad	4514	21.6
فاطمة fa:ṭma	1431	12.6	احمدʔħmd	2329	11

Table 1 shows the percentage of names from the most 20 frequently used names by sex for this particular period. One of these names which is nidha:l نضال consider a neutral name but unfortunately the percentage of this name for newborn female is unexcited.

© 2016 Global Journals Inc. (US)

During this period of time, there were appeared what is called Compound Names which are personal names consisting of more than one name. According to Abd-el-Jawad (1986), classified the compound names into three structural patterns. Firstly, /ʕbd/ عبد servant of God (Allah) plus any of the attributes of God (Allah), of which there are potentially ninety-nine names, such as, /ʕbd eraħjm/عبد الرحيم. Secondly, a verbal noun expressing a positive attribute plus /adjn/ الدين religion, such as, /ṣala:ħidjn/ صالح الدين. Thirdly, a verbal noun expressing blessing and thanking of God (Allah) plus Allah, such as, /ħamdullah/الحمدلله. But compound names, as meant here, are those which consist of the name of the Prophets Muhammad, peace be upon him, plus another common name, such as, /muħamad ʕli:/محمد علي, /muħamad ħasan/محمد حسن and /muħamad jwsif/محمد يوسف .

The political and atmospheric conditions and circumstances, which has happened in this period, impacted the society's daily life, thinking of wars, battles and Jihad. Consequently, the choice of newborn names for male and female were also affected.

iii. *In Relation to Social Factors*

During this period, people lived a simple life, based on religious and social values, where there was little or in some cases outside influence. This can be shown by the social heritage of naming as sons were named after their grandfather and daughters after their grandmothers. The researcher explain that by consider this phenomenon as a tradition in the society which keep the name live not to die or the bearer of such a name is a distinguish person in the family.

b) The Choice of Newborn Names During the Period between the 80s and the 90s

As a result of the process of urbanization and modernization in all aspects of life, including naming, there is a strong tendency to abandon traditional, long, and compound names in favour of modern names which are described by many parents as being short, soft, and easy to write and pronounce, such as, /rana:/ رنا, /rjm/ ريم , and /nu:r/نور .

Based on the analysis of gathered data, the researcher was able to identify the most frequent names chosen during the period between the 80s and 90s.

i. *In Relation to Historical Events*

Due to the Iran–Iraq War, also known as the First Persian Gulf War, the names /ṣadam/صدام which is the name of the late president of Iraq , we noticed that in this period the frequent of chosen the name was 1056 out of 16342 and /ʕudy/ عدي became the most frequently chosen names, as people became affected by the political events of these times, and this is due to the historical relations between the two counties, Jordan and Iraq.

ii. *In Relation to Religion*

In relation to religion, the majority of Jordanian people are Muslims so they mainly choose Arabic-Islamic names. In this period of time, the names chosen still carried the Arabic- Islamic features, such as the female names /fa:ṭma/ فاطمة, /muħamad/ محمد, /ʔema:n/ايمان , /mrjm/مريم and /ʔħmd/أحمد .

iii. *In Relation to Social Factors*

During the 80s and the 90s, the names identified by the researcher, were more authenticity such as the male names, /jazn/يزن , /laiθ/ ليث , /ra:mj/ رامي , and female names /ħna:n/حنان, /raʃa:/رشا, and /nu:r/ نور. This can be analyzed to indicate that the change in the daily lifestyle of the Jordanian people were affected by the appearance of satellite channels podcasting from all around the world, carrying with them not only new cultures but also new thoughts and ideologies, influencing the change in the choice of newborn names, for both male and female.

Also, in these periods of time, compound names were still used whether how Abd-el-Jawad classified them, mentioned and/or noticed above or as the researcher classified previously. The following table shows the most frequent names chosen for this period of time. The names was arranged in the table randomly.

Table 2 : Names and frequency during the 80s- 90s

Female names		Name Frequency	Rate%	Male names		Name Frequency	Rate%
ħna:	حنان	1753	9.0	Jazn	يزن	598	1.7
raʃa:	رشا	1445	7.4	laiθ	ليث	583	1.6
nu:r	نور	1172	6.5	ra:mj	رامي	627	1.8
Fa:ṭima	فاطمة	1340	6.9	ṣadam	صدام	1056	3.1
ʔema:n	ايمان	1753	7.4	ʕudy	عدي	1458	4.2
mrjm	مريم	733	3.7	muħamad	محمد	10126	29.4
ʕa:ʔʃa	عائشة	186	1	ʔħmd	احمد	5947	17.3

© 2016 Global Journals Inc. (US)

It is noticeable that female newborn and male newborn names were fitting, right, or true for persons during this period of time. The names are associated with the increase of modernization and urbanization of Jordanian people.

c) *The Choice of Newborn Names During the Period between the 2000 to the 2015*

i. *In Relation to Historical Events*

In the relation to the historical factors of this period, there were no noticeable changes in the choice of names for newborn males and females, even though these times were full of historical events, such as the Arab Spring, which mainly included all the neighboring countries.

However, in the early 2015, due to the recent "political" events in the Middle East, and moreover, after the brutal murder of the Jordanian pilot "Moath Al-Kasasbeh" معاذ الكساسبة, many Jordanian newborns have been given the name /mua:ð/ معاذ. This can be taken as a symbol of loyalty and unity of the Jordanian people.

ii. *In Relation to Religion*

In relation to religion, the researcher has noticed that newborn male names remained to carry the Arab-Islamic feature with only a few foreign names, whereas, the Arab-Islamic features of female newborn names had decreased to the state where they are no longer used, or only a small percentage remains present. The researcher predict that the newborn female names which carry an Arabic-Islamic features will eventually disappear or at least become even less used in the future years to come.

iii. *In Relation To Social Factors*

In the last two decades, Jordan, as many other communities in the Middle East, has witnessed and experienced tremendous, modernization, urbanization, globalization, and the Internet. Due to these revolutions, the world has become as some say "a small village", therefore, different societies and cultures have intertwined, borrowing and exchange not only terms and expressions, but also personal names. The consequence of this international communication is the appearance of new and foreign names bestowed upon male and female newborn names. So, it has become natural to hear of an Arab girl called /ʔlma:/ الما or /sjljn/ سيلين or an Arab boy called /kina:n/كنان or /ʔmjr/امير. The choice of such names seems to be due to that foreign names are easy to pronounce, spell, and recall. Also, mass media during these particular periods play a significant role in developing this phenomenon, as Jordanian, like other nations, watch foreign programs and series on TV influencing lives.

Based on the data gathered, the researcher noticed, that during the year 2000 to the 2015, that there was a large increase in the number of newborns given foreign names, for instance, the female names /ʔeljn/ ايلين, /twljn/تولين, and /rjta:l/ريتال and the male names /rja:n/ ريان ,and /ʤa:d/ داج . This phenomenon of newborns with foreign names is thought to have arisen due to the large number of Turkish and Western TV series and movies. A foreign newborn male name and newborn female name signify modernity, prestige and pride. To the opposite, this phenomenon of personal foreign naming could be a strategy of breaking away from traditional values. The following table shows the frequently chosen names for this period of time.

Table 3 : Names and frequency during 2000-2014

Female names		Name Frequency	Rate	Male names		Name Frequency	Rate
ʔeljn	إلين	893	1.1	rja:n	ريان	1638	1.7
lja:n	ليان	2129	2.4	ʤa:d	جاد	534	0.6
rjta:l	ريتال	1128	1.3	ʔmjr	أمير	2488	2.7
mja:r	ميار	1197	1.4	ʕbd alrħma:n	عبد الرحمن	3467	3.7
fa:ṭma	فاطمة	432	0.5	Muħamad	محمد	12859	13.7
ʕa:ʔʃa	عائشة	112	0.1	ʔħmd	أحمد	4781	4.9

As shown in Table 3, these particular names were chosen, and arranged randomly, as they are the most frequently given birth names for both genders during this period as the result of the globalization and modernization.

Data analysis reveals a sex differentiation in naming patterns. The results seem to support one of the most important findings about sex-based language variation which emerged from sociolinguistic studies in the industrialized western world. These studies prove that women tend to be on the average to use more higher-status and prestigious linguistic change taking place in the direction of socially prestigious linguistic forms, and if new linguistic forms are introduced, women will be ahead of men (labov 1972, Trudgill 1974, and Chambers and Trudgill 1980).

© 2016 Global Journals Inc. (US)

Since a name is a symbol of identification, it can be also a symbol of social prestige. For females, a name is consider as a part of her beauty; therefore, parents choose an attractive name for their daughters, so they search carefully to find an attractive one which is beautiful both in sound and meaning. On the other hand, for males, a name is consider as a part of his personality; therefore, when parents choose a name for their son they select traditional and old names which once were very common. It could be said as in the Arab proverb "To every person his name is fortune". The Prophet, peace be upon him, changed the names of many of his companions to carry the best and noblest meanings, and he changed Amr Bin Hisham's nickname from Aba Alhikmato Abu Jahl for his insistence on infidelity.

During the study the researcher has noticed that compound names are no longer used as the use of these kinds of names have been discouraged by the government in this period of time.

Table 4 : Arabic- Islamic Names

Period	Female name		Male names	
	Fa:ṭma	ʕa:ʔʃa	Muhamad	ʔhmd
70s	1431	401	5559	2575
80s-90s	1318	186	8813	4707
2000-2015	432	112	15642	4781

As shown in Table 4, the main difference between male and female is the disappearance of religious female names such as /fa:ṭma/ فطمة the frequency of the name in the 2015 compared to the previous generations. The researcher believes that in the forthcoming days with the rapid increase in the use of English and foreign names this will further exist to effect the choice of names as, the newborn male names carrying an Arabic-Islamic features and which indicate qualities and attribution of manhood, courage, nobility and generosity will die out as well as newborn female names which have started to die since the beginning of the process of modernization and globalization.

These particular names, as shown in table 4, have been chosen for their religious significance, as these names will never die out or become unused. Religious feelings reflect on the lifestyle, pattern of thinking, and consequently the linguistic behavior including naming. As noticed, people during the seventies, people refused the influence of the Western culture and decided to keep the choice of Arab Islamic prestigious names but with the passage of time and with the emergence of urbanization or modernization and technology people seemed to accept the entry and the influence of the Western culture and then the acceptance of its effect, also in the choice of names.

Throughout this study, the researcher noticed that the first two generations the parents derive the names of females from the names of males, such as the female name /ʃa:dja/ شاديه from the male name /ʃa:dj/ شادي, /fa:dja/ فاديه from /fa:dj/فادي and /fatħja/ فتحيه from the name /fatħi/ فتحي and so on. But this phenomenon has almost disappeared nowadays.

d) Phonetic Change of Sound Pattern Personal Names through Times

While analysing the gathered data, the researcher observed that there are a number of phonetic changes in sound pattern of personal names, especially in female names.

The data shows that the phonetic sounds of personal names have changed throughout time. In comparing and contrasting personal names from the 70s with those from the 2014-2015, we can see that in the past names used to be heavy and consist of dark sounds, such as the dark /l/ sound. As in, the female names /xawla/ خولة , /nawa:l/ نوال and /xolwd/ خولد, whereas the 2015 names seem to be more lighter and more rhythmic, such as the female names, /ʔeljn/ ايلين, /ljn/ لين and /ja:ra:/ اراي . This can be mainly seen in female names more than male names. This phenomenon formed maybe as a result of the political chaos surrounding the community, affecting clearly and concretely on the choice of names, making it possible to say that the political chaos formed a mess in the selected names.

e) How the Names are Chosen within the Same Family

As another interesting fact about naming newborns, throughout the study the researcher observed that there was, in some families, a trend of choosing similar names for siblings that making their family distinguished from others, many strategies may be used: The first one, is by the use of similar sound for all children, such as /su:ha/ سهى, /nu:ha/نهى, and /maha/ مها . The second, is by the use of the same beginning sound, such as /xulwd/ خولد and /xalid/ خالد. The third, is by the use of the similar ending sound to name all children such as /alla:ʔ/ ألاء, /baha:ʔ/ بهاء, /wafa:ʔ/ وفاء and /ʕala:ʔ/ علاء.

These strategies can be visible in many families in Jordan, a possible justification for the use of these strategies is that it makes it more easier for the parents to call their children, as some find it joyful for if they call one child, all may answer as they all have the same

© 2016 Global Journals Inc. (US)

sounds in their names. Another justification is that some parents find it more musical to name all their children using the same rhythm. And we can say that parents subconsciously substitute similar sounding sibling names more often than dissimilar names.

IV. Summary and Conclusion

This present study, which arose from a sociolinguistic interest, is an identification of how the choice of names for newborn children in Jordan has changed throughout the eras, as documented in the official database at the Department of civil status and by the Department of Statistics. It also aimed at identifying the reasons behind the difference in choice of names given to newborn children in Jordan, in relation to the social, religious, and historical events, which occurred during the eras in question.

The researcher came up with the following conclusions. Generally, the change in the choice of names for newborns is quite evident. Especially, in relation to the events, circumstances, and political conditions affecting the Jordanian society during the times, this can be most evident during the 1970s all through to the early year of 2015. With the evidence of the choice of mostly 'Jedahi' names in the 70s, and the choice of naming newborns after well known politicians during the 80s to 90s, due to the wars and political events which happened during this decade, as mentioned previously. Such trend seemed to fade during the years 2000 to 2014, despite the political events present at such time, the Arab Spring. However, at the beginning of the year 2015, it was evident once more with the obvious increase in the choice of the name /mua:ð/عاذ after the Martyr Jordanian pilot 'Moath Al-kasasbeh'.

The political and atmospheric conditions and circumstances, which have happened during these periods of time, have had a great impact on the society's daily life, resulting in their thinking of wars, battles and Jihad. Consequently, the choice of newborn names for male and female were also affected.

With regard to religion, the findings show that the choice of Arabic-Islamic names is evident during the 70s, for both males and females, as newborns were named after Prophets, messengers, and famous religious figures. During the 80s to 90s, male's names still seemed to carry Arabic-Islamic features, whereas, a minor change in the choice of female names was visible, as they became lighter and more modern. This became more evident during the years 2000 to 2015, as the Arabic-Islamic features of female names decreased and the appearance of more names carrying no religious features what so ever. It was predicted that the choice of female Arabic-Islamic names will eventually become rare and therefore less used in the future years to come.

Religious feelings reflect on the lifestyle, pattern of thinking, and consequently the linguistic behavior including naming. As noticed, people during the 70s, refused the influence of the western culture and decided to keep the choice of Arabic-Islamic prestigious names. On the other hand, as time proceeded with the emergence of urbanization/modernization and technical development, people seemed to accept the entry and the influence of the Western culture and then the acceptance of its effect, not only on their lifestyle, but also on their choice of names.

With regard to the social features affecting the choice of names for newborns in Jordan, it was quite evident during the 70s that family values played a great role in the choice of names, as children were named after either their grandfather or grandmother or even after an uncle or an aunt. This comes to show how close and intermit families were during this time. When coming to the 80s and 90s, it was evident that the choice of names for newborns was affected by technical development, moreover, the home invasion by satellite channels resulted in the enlightenment of the Jordanian people to different cultures and ideologies, and even more, resulting in the choice of foreign and more modern names for their newborns. During the 2000 to 2015, a large increase in the number of foreign names given to newborns, especially females was evidently visible. This is mostly due to the Internet and the increase of Turkish and Western television series, as people have become influenced by these TV programmes that they began naming their children after certain characters; even though these names are of Arabic origin, they have not been used in the same way as after the appearance of dubbed TV series. This can be seen as a strategy for steering away from traditional values towards modernization.

Accordingly, for a person, his/her name becomes the property of his/her personality. Like every aspect of society, the name, which is a symbolic system of identification, is usually "historically constructed, socially maintained, and based on shared assumption of a particular community" (Akinnaso 1980:227). In the present days, in Jordan, the bestowal of a newborn's name is the parents' duty.

Since a name is a symbol of identification, it can be also a symbol of social prestige. For females, a name is consider as a part of her beauty; therefore, parents choose an attractive name for their daughters, so they search carefully to find an attractive one which is beautiful both in sound and meaning. On the other hand, for males, a name is consider as a part of his personality; therefore, when parents choose a name for their son they select traditional and old names which once were very common.

Furthermore, while reviewing the gathered data, more evidence of change in the choice of newborn names surfaced, either being in the structure, meaning and/or gender.

© 2016 Global Journals Inc. (US)

Firstly, it was noticed that during the 70s to the late 80s some female names were derived from names that of males, such as the female name /ʃa:dja/ شاديّة from the male name /ʃa:dj/ شادي , /fa:dja/ فاديّة from the male name /fa:dj/,فادي and /fatħja/ فتحيّة from the male name /fatħj/ فتحي and so on. But as coming to the 90s, this phenomenon has almost disappear.

Secondly, the phonetic differentiation of personal names throughout the eras; It was observed that there is a number of phonetic changes in the sounds of personal names, especially in female names.

Findings show that the phonetic sounds of personal names have changed throughout time. In comparing personal names from the 70s with those from the 2014-2015, in the past names used to be heavy and consist of dark sounds, such as the dark /l/ sound, (e,g., the female names /xawla/ خولة , /nawa:l/ نوال and /xulwd/ (دولخ, whereas, in 2014-2015 names seem to be more lighter, rhythmic, shorter and easier to pronounce and spell, such as the female names, /ʔeljn/ايلين , /ljn/ لين and /ja:ra:/ اراي . This can be mainly seen in female names more than in male names.

Thirdly, naming newborns within the same family, i.e., brothers and sisters. As another interesting fact about naming newborns, it was observed that in some families there was a trend of choosing similar names for siblings. In order to make their family distinguished from others.

A possible justification for the use of this is that it makes it easier for parents to call their children, as some find it joyful as if they call one child, all the other children may answer as they all have the same sounds in their names, making it hard to distinguish which name is being call. Another justification is that some parents find it more musical to name all their children using the same rhythm ending. And it can be said that parents subconsciously substitute similar sounding sibling names more often than dissimilar names.

Fourthly, during the 70s there were as called 'Compound Names' which are personal names consisting of more than one name. According to Abd-el-Jawad (1986), classified the compound names into three structural patterns:

> /3bd/ عبد servant of God (Allah) plus any of the attributes of God (Allah), of which there are potentially ninety-nine names, such as, /ʕbd eraħjm/عبدالرحيم .
>
> a verbal noun expressing a positive attribute plus /adjn/الدين religion, such as, /ṣala:ħidjn/صالح الدين.
>
> a verbal noun expressing blessing and thanking of God (Allah) plus Allah, such as, /ħamdu:llah/ الحمدلله.

But compound names, as meant here, are those which consist of the name of the Prophets Muhammad, peace be upon him, plus another common name, such as, /muħamad ʕlj/ محمد علي, /muħamad ħasan/ محمد حسن and /Muħamad jwsif/ محمد يوسف. Mainly the Jordanian government has discouraged the use of such compound names, as much confusion occurs in result of such names. As it is hard to distinguish if the second name is part of a compound personal name or that if it is the person's father's name.

Finally, as a general aspect noticed to be present in all periods in question, that there are some female names written/spelled in more than one way, such as the name /ra:nja/ ((رانيه ، رانيه ، رانيا which is written with three ending variations"ـه" or "ـة" or "ـا", all carrying the same meaning and pronunciation, but different spelling. This phenomenon may be related to the non-standardized spelling of such names by the community.

In conclusion, personal names are useful tools for reference as they provide information on various events in the life of an individual, the family or the society.

References Références Referencias

- English References

1. Abd-el-Jawad, H. (1986). A linguistic and sociocultural study of personal names in Jordan. Anthropological Linguistics 28(1), 80-94.
2. Adjah, Olive Akpebu. (2011). What Is in a Name? Ghanaian Personal Names as Information Sources. African Research & Documentation, No. 117, September 1, 2011. Web. 3 Jan 2015.
3. Agyekum, Kofi. (2006). The Sociolinguistic of Akan Personal Names. Nordic Journal of African Studies 15(2): 206–235.
4. Akinnaso, F. (1980). "The Sociolinguistic Basic of Yoruba Personal Names". Anthropological Linguistics, Vol: 22. NO.7. PP. 275-305.
5. Alokaily, Jalal. (2005). "A Semantico – cultural Study of Place Names in Jordan". MA Thesis, Yarmouk University, Irbid, Jordan.
6. Al-Shahi, A. (1988). "Some Personal Names and Nicknames among the Rivera in People of Northern Sudan". Journal of the Anthropological Society of Oxford19 (2):131-139.
7. Al-Tahat, Abdelrazzaq. (2014). "A Socio- Semantic Study of Family Names in Jordan". MA Thesis, Yarmouk University, Irbid, Jordan.
8. Al-Zumor, Abdul Wahed. (2009). "A Socio-Cultural And Linguistic Analysis Of Yemeni Arabic Personal Names". GEMA Online Journal of Language Studies.15Volume (2).
9. Babycenter.com. (2014). Top 10 baby names around the world 2014. BabyCenter, L.L.C. 2014. Web. Jan 3, 2015.
10. Bader, Yousef. and M. Salih. (1997). "A Diachronic Analysis of Personal Names in Jordan". Interface: Journal of Applied Linguistics 12(1): 3-13.
11. Crystal, David. (1995). Phoneasthetically Speaking". In English Today, 11:2, pp.8-12.

© 2016 Global Journals Inc. (US)

11. Cutler, A. And others (1990) "Elizabeth and John: sound pattern of men's and women's names". In Journal of Linguistics, 26:2, pp. 471-482.
12. Encyclopedia Britannica. (1926). London: The Encyclopedia Britannica Inc., Volume 19.
13. Jayaraman, Raja. (2005). "Personal Identity in a Globalized World: Cultural Roots of Hindu Personal Names and Surnames". The Journal of Popular Culture, Vol. 38, No. 3, 2005.
14. Kimenyi, A. (1978). "Aspect of Naming in Kinyarwanda". Anthropological linguistics 20(1): 258-271.
15. Kwinika, Savious. (2010). South Africa World Cup baby names: Moms choose Fifa, Bafana, and Soccer City. Christian Science Monitor, 08827729, 6/22/2010. Database: EBSCOhost.
16. Liseli, A.Fitzpatrick,B.A.(2012). "African Names and Naming Practice. The Impact Slavery and European Domination Had on the African Psyche Identity and Protest''. MA Thesis, Ohio State University, USA.
17. Mehrabian, Albert, and Marlena. Piercy. (1993). "Differences in Positive and Negative Connotations of Nicknames and Given Names". Journal of Social Psychology 133 (5): 737-739.
18. Mphande, lupenge. (2006). "Naming and Linguistic Africanism in African American Cultural". Ohio State University. Available Online: Web. March 11, 2014.
19. Pappas, Stephanie. (2013). Baby Names Reveal Parents' Political Ideology. Live Science Contributor. June 05, 2013. Web. Jan 3, 2015.
20. Salahuddin, Ahmad. (1999). "A Dictionary of Muslim Names". United Kingdom. London.
    Salih, M., and Y. Bader. "Personal Names of Jordanian Arab Christians: A Sociology of Language".140:29-43.
21. Salih, M. (1999). "Foreign Personal Name-Giving to Arab Girls: Attitudes and Factors Among Arab Youths". Abhath Al-Yarmouk: Lit.&Ling. Vol 17, No. 2, 1999, pp-28.
22. Trends in Naming Babies. (2001).The Christian Science Monitor. 1/10/2001, Vol. 93 Issue 32, p14. 0p. Database: EBSCOhost.
23. Wardat, M. (1997). "Nicknaming in Jordanian Arabic: A Sociolinguistic Perspective". Interface: Journal of Applied Linguistics12 (1): 45-58.
24. Yassin, M. (1978). "Personal Names of Address in Kuwaiti Arabic". Anthropological Linguistics. 20 (2):53-63.
25. Zuraiq, M. (1999). "Linguistic and non-Linguistic Factors Influencing Name Giving in Jordan". MA Thesis, Yarmouk University, Irbid, Jordan.
26. http://www.alwakeelnews.com/index.php?page=article&id=126427#.VNMKzp2Ude8
27. http://www.assakina.com/mohadrat/16603.html#ixzz3tI8ByXUl
28. http://www.oxforddictionaries.com/definition/English/name

- Arabic References

المراجع العربية
القرآن الكريم
رأفت، سويلم. (2001) تربية الأطفال في الإسلام .

© 2016 Global Journals Inc. (US)

This page is intentionally left blank

© 2016 Global Journals Inc. (US)

Global Journal of human-social science: G
Linguistics & Education
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-460X & Print ISSN: 0975-587X

Agglomerative Hierarchical Clustering: An Introduction to Essentials. (1) Proximity Coefficients and Creation of a Vector-Distance Matrix and (2) Construction of the Hierarchical Tree and a Selection of Methods

By Refat Aljumily

University of Newcastle, United Kingdom

Abstract- The article is on a particular type of cluster analysis, agglomerative hierarchical analysis, and is a series of four main parts. The first part deals with proximity coefficients and the creation of a vector-distance matrix. The second part deals with the construction of the hierarchical tree and introduces a selection of clustering methods. The third deals with a variety of ways to transform data prior to agglomerative cluster analysis. The fourth deals with deals with measures and methods of cluster validity. The fifth and final part deals with hypothesis generation. The present article covers the first and second partsonly. It explains how agglomerative cluster analysis works by implementing it in a data matrix step by step.

Keywords: proximity, metric space, vector space, (non) euclidean space, symmetric matrix, agglomeration, centroid, sum of squares, median.

GJHSS-G Classification : FOR Code: 139999

Strictly as per the compliance and regulations of:

© 2016. Refat Aljumily. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http:// creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Agglomerative Hierarchical Clustering: An Introduction to Essentials. (1) Proximity Coefficients and Creation of a Vector-Distance Matrix and (2) Construction of the Hierarchical Tree and a Selection of Methods

Refat Aljumily

Abstract- The article is on a particular type of cluster analysis, agglomerative hierarchical analysis, and is a series of four main parts. The first part deals with proximity coefficients and the creation of a vector-distance matrix. The second part deals with the construction of the hierarchical tree and introduces a selection of clustering methods. The third deals with a variety of ways to transform data prior to agglomerative cluster analysis. The fourth deals with deals with measures and methods of cluster validity. The fifth and final part deals with hypothesis generation. The present article covers the first and second partsonly. It explains how agglomerative cluster analysis works by implementing it in a data matrix step by step. Different types of agglomerative hierarchical clustering methods are applied on purposely-made data matrix so different types of cluster structures are made from that same dataset. The last three parts will be covered in the next publication(s).There are many articles, tutorials, and books on this subject. The article has two main objectives: (1) to keep the discussion short and easy to understand by (hopefully) any reader and (2) to develop the motivation for using agglomerative hierarchical clustering to analyse any high-dimensional data of interest with respect to some research question.

Keywords: proximity, metric space, vector space, (non) euclidean space, symmetric matrix, agglomeration, centroid, sum of squares, median.

I. Introduction

Agglomerative Hierarchical Cluster Analysis, abbreviated (AHCA), is a particular type of cluster analysis and is a useful multivariate exploratory technique that has found application in different research fields such as data mining, social sciences, biology, information retrieval, statistics, pattern recognition, ecology and psychology. Agglomerative Hierarchical Cluster Analysis is not a single method but rather a family of different but related computational methods that make no a priori assumptions about the structure of data. Agglomerative Hierarchical Analysis methods try to discover structured interrelationships among data vectors that might be interesting in relation to a research purpose. More specifically, all the methods of the family try to identify and graphical display of structure in data when data is too large either in terms of the number of variables or of the number of objects described, or both, for it to be readily interpretable by direct inspection. Agglomerative Hierarchical Analysis methods generate hierarchically ordered clusters and represent proximity structure among objects in high-dimensional space not as a spatial cluster but as a constituency tree or dendrogram. All the methods work by grouping a set of objects in the domain of interest into distinct clusters according to how relatively similar/dissimilar those objects are in terms of the variables that describe them. Each object is described by a set of variables. Any two objects will be more or less similar/dissimilar on the basis of some definition of proximity between them.

This article is in four main parts. The first part gives a general description of agglomerative hierarchical cluster analysis and proposes an interpretation of the result related to it. The second part first provides some relevant mathematical concepts that will be used in agglomerative hierarchical clustering: cluster, metric space, vector space, and proximity matrix, and then goes into the detail of how proximity among pairs of vectors is measured and how a cluster tree is built. The third part shows twelve different varieties of agglomerative hierarchical analysis and applies them to a data matrix M. The final part concludes the discussion.

a) *Agglomerative Hierarchical Cluster Analysis (AHCA) and interpretation*

AHCA is known as a bottom-up or alternatively left to right approach. This approach is the more often used and also better covered in the relevant textbooks, e.g., [1], [2], and [3]. This is probably because AHCA provides more information than the other methods in that they not only identify the main clusters, but also their constituency relations relative to one another as

Author: School of English Literature, Language and Linguistics, University of Newcastle, Newcastle Upon Tyne, Tyne and Wear NE1 7RU, UK. e-mail: refat.aljumily@newcastle.ac.uk

© 2016 Global Journals Inc. (US)

well as their internal structures. The result of the utilization of AHCA is shown by a diagram called a 'constituency tree' or 'dendrogram', which groups together related data vectors based on the relativities of proximity among all pairs of data vectors. Figure/1shows the result from the application of AHCA to eight data vectors.

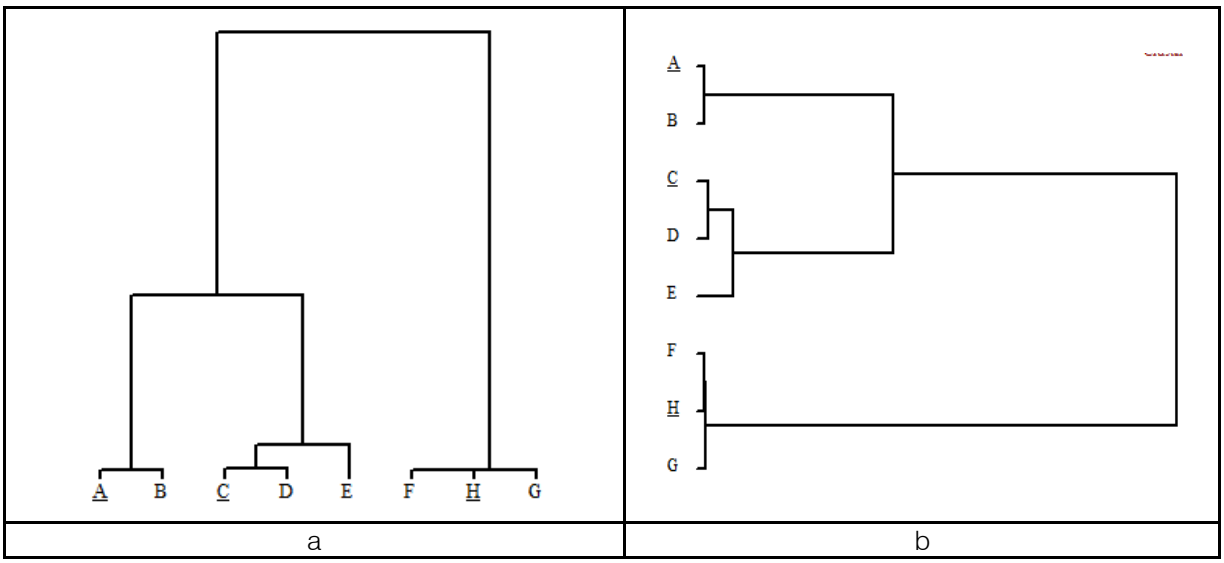


Figure 1 : (a) Vertical view of bottom-up tree and (b) horizontal view of left to right tree of five data items

Figure/1shows the cluster structure of eight data vectors as a hierarchical dendrogram. To interpret the dendrogram correctly one has to understand how it is constructed, so a short intuitive account is given here; technical details are given later in the course of the discussion. The dendrogram in this figure can be viewed in different ways, that is, either vertically (a) or horizontally (b). In it the letters at the leaves are labels for the vectors in the dataset: "A" is the first vector, "B" the second, and so on. These labels are agglomerated into clusters in a sequence of steps. AHCA treats each data vector as a single cluster on its own and then sequentially agglomerate pairs of clusters until all clusters have been agglomerated into a single larger cluster that contain all data vectors. The links included in the hierarchy represent the constituency structure for the entire dataset: vector "A" and vector "B" constitute a cluster (A B), vector "C" and vector "D" constitute a cluster (C D), which itself combines with vector "E" so constitutes a cluster ((C D) E) that are combined together with (A B) to form an even higher-level cluster ((A B) ((C D) E)), and so on. The lengths or heights of the links represent degrees of closeness: the shorter the link, the more similar the clusters. This is reflected in the cluster tree by the relative lengths of these links by the constituency structure of the proximity relations among, for example, vectors (A B)and vectors (F H) or vector (G). The longest (vertical/horizontal) lines at the top or right of the dendrogram separate the vectors into three main groups. The dendrogram represents vector proximity in n-dimensional space. For example, vector "F" and vector "H" are very close in the data space, and this pair is close to vector "G".

II. Space Concepts

a) Cluster Definitions

From cluster analysis viewpoint, the power of human eye or brain can recognize structures that are contained in data by perceiving any clusters in it, despite the fact that the clusters may vary somewhat in different viewpoints, in many different sizes and shapes or even when they are interpreted or understood. To accept such a view we have to understand what a cluster is. Indeed, humans can detect patterns or connections in any surrounding environment and can distinguish between them, and clusters are a kind of pattern. In a countryside position, for example, we can see clusters of trees, or farm buildings, of sheep. In any clear night we can see in the sky clusters of stars. And, closer to current interests, anyone looking at a data plot immediately sees any clusters that might be present. Looking at the data plotted in the two-dimensional space below, on the basis of our innate pattern recognition capability and without recourse to any obvious definition of the cluster, we can see that in figure/2a there is a random cloud of points with no clear structure emerging behind the data, and that in figure/2b there are some local areas of concentrations of points, but these are not explicitly defined. By contrast, we can clearly see that in figures/2c and 2d there is a clear structure: figure/2c shows three clusters of equal size, whereas figure/2d shows two clusters of unequal size, the smaller of which is in the upper-left part of the plot and the larger one in the lower part.

© 2016 Global Journals Inc. (US)

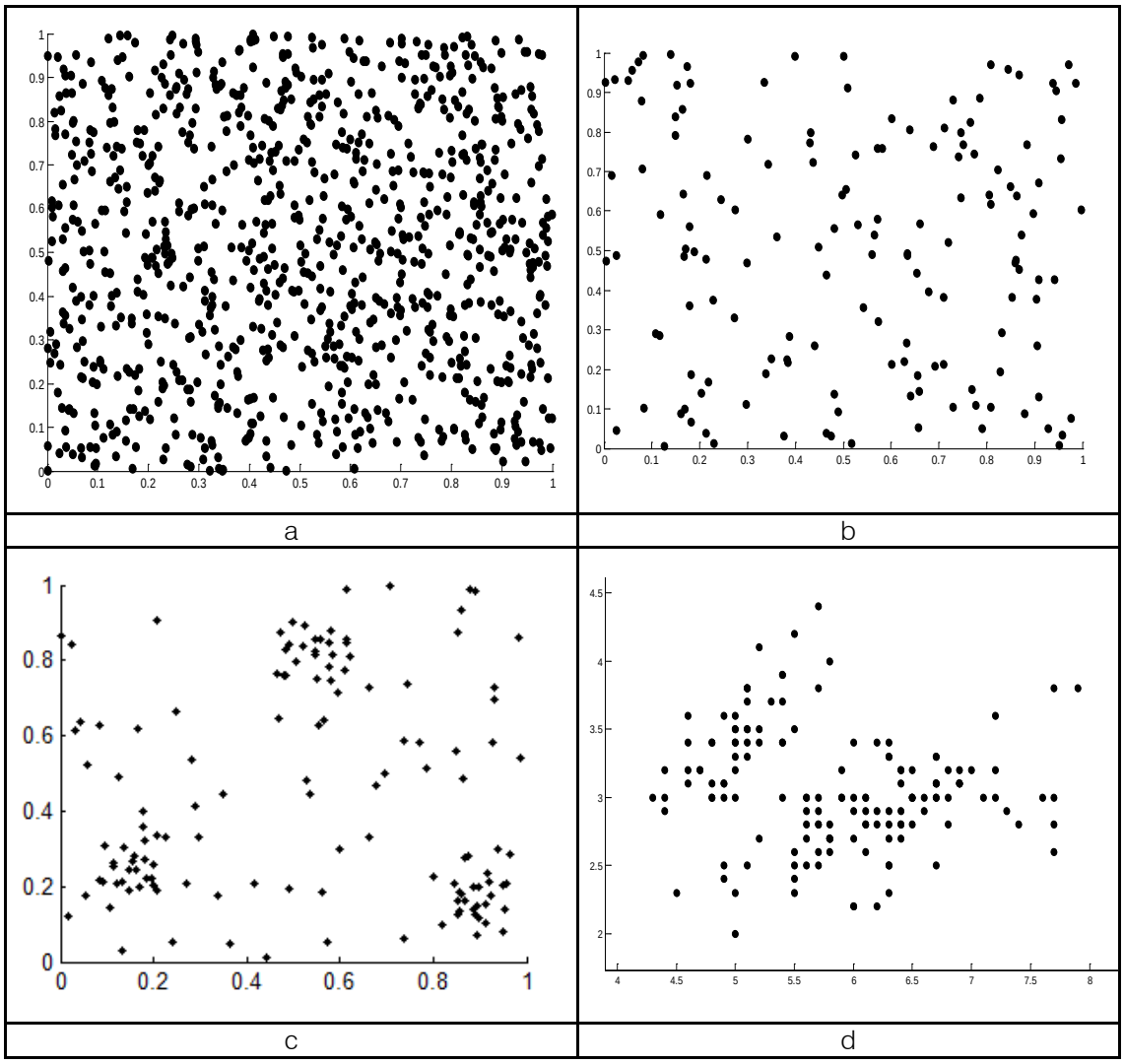


Figure 2 : A random scatter of points in two dimensions

The term cluster, however, does not have a precise definition, but there are some working definitions of what a cluster is that are commonly used. Three of them are given by [4] and [5]. They are:

- "A cluster is a set of entities which are alike, and entities from different clusters are not alike";
- "A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it";
- "Clusters may be described as connected regions of multi-dimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points".

The first definition of a cluster is a very general one and is best described as a similarity-based cluster definition. It assumes that objects are similar to each other within the same cluster and dissimilar to objects in different clusters. The second introduces the distance view of similarity and is best described as a distance-based cluster definition. It assumes that the similarity or dissimilarity between data vectors can be measured on the basis of the distance between them. The third definition of a cluster introduces density view of similarity and is best described as a density-based cluster definition. It assumes that each cluster is representing a given region that has its own demand distribution which symbolizes the data vectors enclosed by that region. This definition is more often used when the clusters are irregular or intertwined, and when noise and outliers are present [6].

Considering these three working definitions, we can see that even if the clusters consist of entities, points, or regions, the data vectors within each cluster are more similar in some respects than are other data vectors outside the clusters. A cluster is therefore a collection of data vectors which are similar between

© 2016 Global Journals Inc. (US)

them and are dissimilar to the data vectors belonging to other clusters. Figure/3 shows a sample data vectors plotted (left) with its corresponding clusters (right) on a two-dimensional scatter plot.

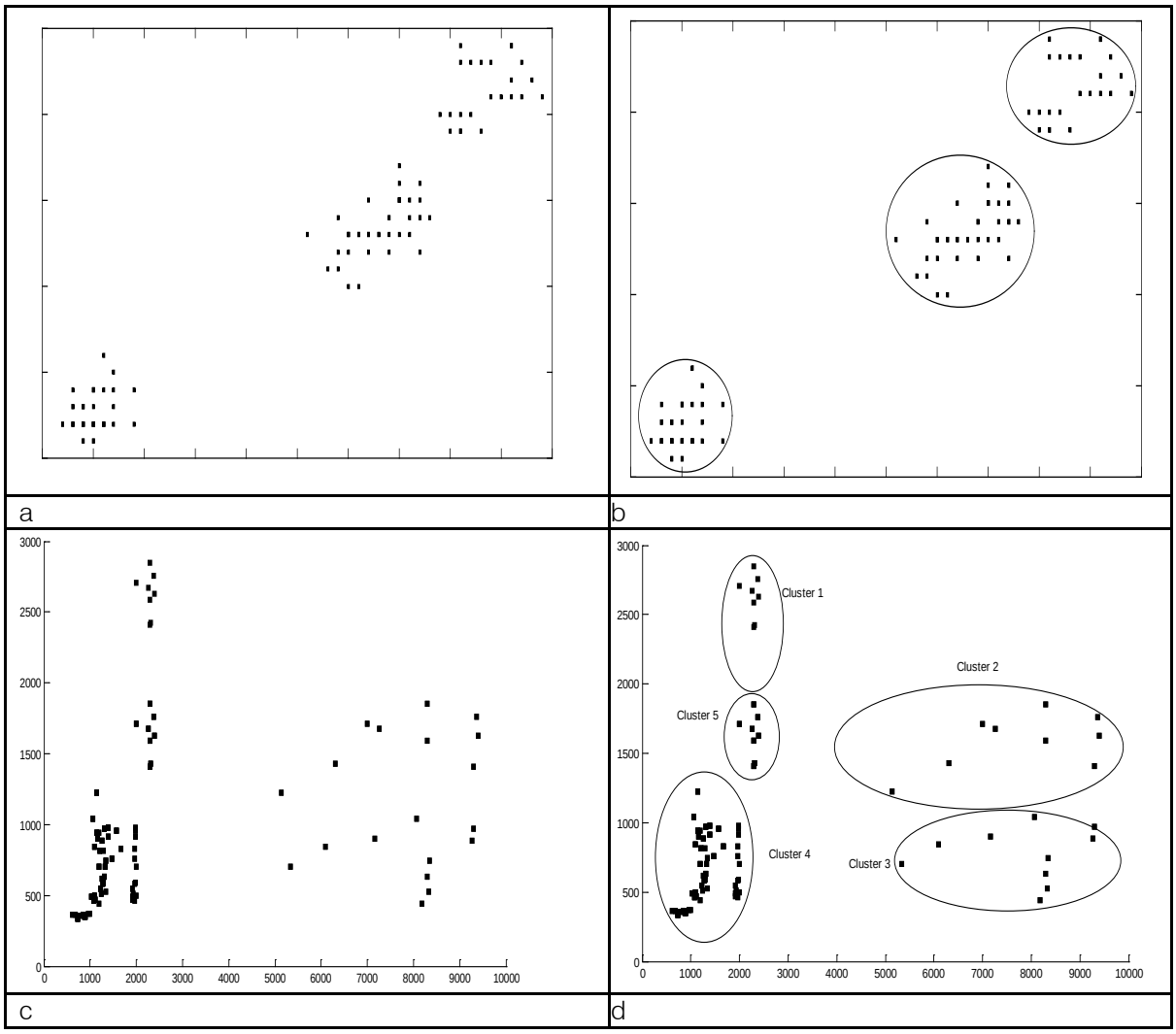


Figure 3 : A scatter of points (left) and its clusters (right) in two dimensions

In Figure/3a-b, the data vectors are clustered into three clear clusters labeled (cluster1, cluster2, cluster3) and in Figure (3c-d), the data vectors are clustered into five clear clusters labeled (cluster1, cluster2, cluster3, cluster4, cluster5) on the basis of some definition of proximity. If anyone is going to attempt an AHCA on data, then he/she should address the issue of what proximity coefficient to use at an early stage.

b) Proximity coefficient

Cluster analysis, by definition, is a process of identifying those data vectors that are similar and of establishing a hierarchical classification relationship among them on the basis of some index of proximity. What we mean by "on the basis of some index of proximity" is to calculate how data vectors plotted as points in multidimensional space are "close to" or "far away from" each other. To do so, we need to know the relative proximity between any two data vectors in different clusters.

A proximity coefficient is either a similarity or distance coefficient between every pair of data vectors in the space. The term proximity is more commonly used to refer to either one of these two coefficients. The term of proximity always suggest the question: proximity with respect to what? Most clustering procedures use pairwise measures of proximity. Two data vectors are close when their distance is small or their similarity is large. The choice of proximity coefficient is a crucial problem in cluster analysis [4]. The choice of which proximity measure to use in the first place is largely a matter of the type of data collected. All clustering information must be built up from the basic data types in the space. The type(s) of data collected in a given study determine the type of clustering analysis used. Most of the clustering algorithms can be applied to only certain kinds of data and some particular measures of

© 2016 Global Journals Inc. (US)

distance/similarity. As Everitt et al. [2] points out different proximity coefficients can and do lead to different cluster solutions, and as such it would be extremely useful to be able to select a proximity coefficient that is in some sense optimal. No reliable selection procedure exists, however. The choice of coefficient in any given application is governed by the nature of the data and by the clustering algorithms that will use it. There are many different types of data that one can collect. The following is a diagram showing some types of data that can be expressed either in terms of numbers or a natural language description.

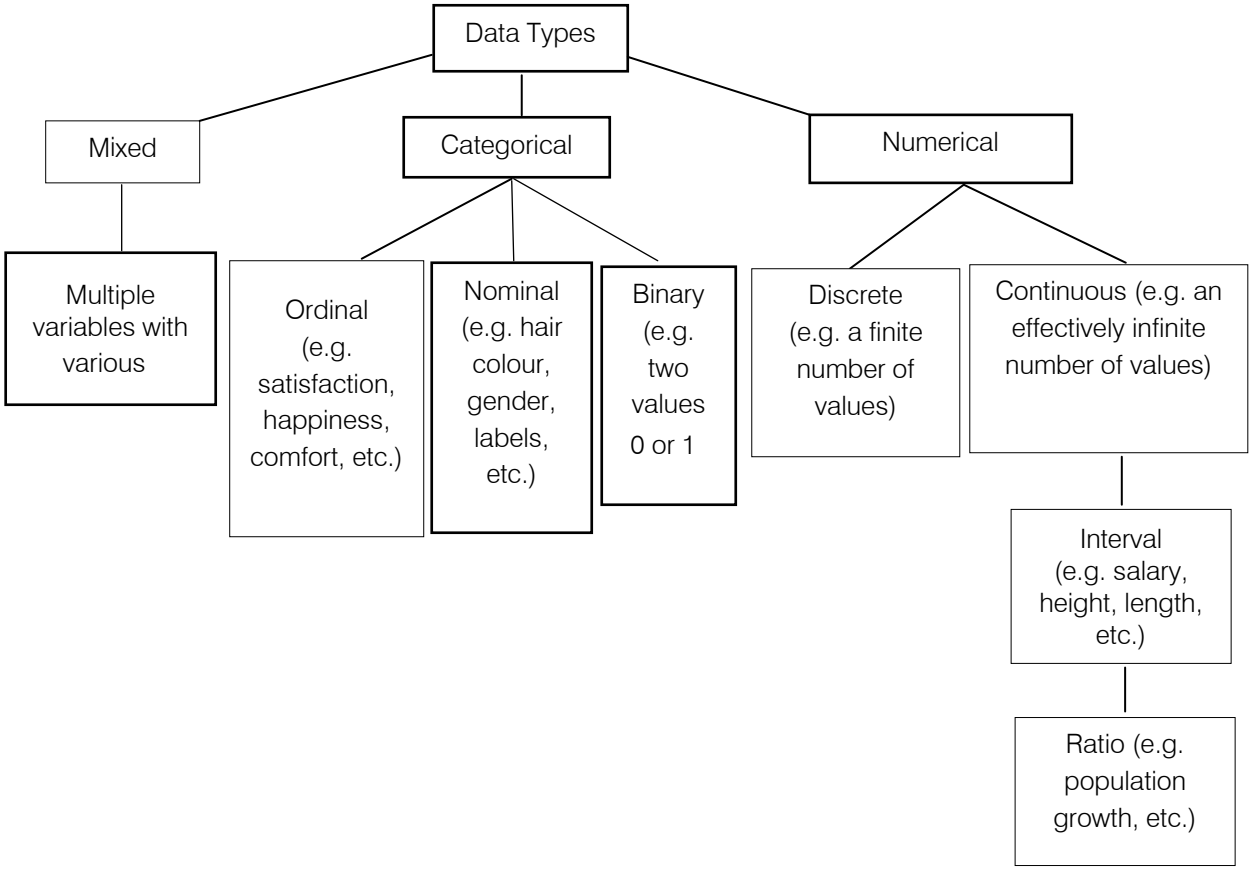


Figure 4 : Types of data

For details of each of the data types see, for example, [7], [8], and [9].

However, proximity between pairs of data vectors can be measured in terms of their correlation, of their similarity coefficient, of the angle between them, or of distance in Euclidean space [2]. With data in which all the variables are categorical, measures of similarity are most generally used. The most commonly used similarity coefficient, at least for binary data is the Jaccard similarity coefficient and is calculated as: $Sij=a/(a+b+c)$. To illustrate, Table/1 gives a matrix of binary variables of dimension 6 x 8.

© 2016 Global Journals Inc. (US)

Table 1 : A 6 x 8 data matrix

	Variables							
	*x*1	*x*2	*x*3	*x*4	*x*5	*x*6	*x*7	*x*8
Vectors A	1	0	0	1	0	0	1	0
B	0	1	0	1	0	0	1	0
C	1	1	1	1	0	0	1	1
D	0	1	1	1	0	0	0	1
E	0	1	1	1	0	0	0	1
F	0	1	0	0	0	0	1	1

Where each row vector is a student and the column vectors are binary tags or states of some student response, e.g. answer to test questions. The state (1) means a variable is present indicating a correct answer in the data vectors and (0) means it is absent indicating an incorrect answer. This data can be summed and placed in a contingency table in the form of the count of the number of the variables in each vector. The first two column data vectors (A) and (B) are worked out and the coefficient of matches among them are shown in Table/2.

Table 2 : A 2 x2 contingency table for the first two vectors in Table (1)

		Vector B	
		1	0
Vector A	1	*a*=2	*b*=1
	0	*c*=1	*d*=4

In this table, the rows represent the presence or absence of a set of *X* variables for a single student {*x*1, *x*2,……*x*8} for the first two row data vectors in Table/1. Cell *a* includes the count of the number of the *X* variables for which the two vectors both have the variable present. Cell *b* represents the number of variables the number of variables for which the first has the variables present and the second does no, and cell *c* includes the number of variables for which the second student has the variable present and the first student does not. Finally, cell *d* includes the count of the number of the *X* variables for which neither student has the variable present.

Now we compute the similarity coefficients between the students based on the coefficient of matches by using Jaccard similarity coefficient. The equation used to calculate the similarity between data vector A and B is the following:

$S_{AB}=a/(a+b+c)= 2/4= 0.500.$

Applying this equation to the other row data vectors:

$S_{AC}=a/(a+b+c)= 3/6= 0.500$

$S_{AD}=a/(a+b+c)= 1/6=0.167$

$S_{EF}=a/(a+b+c)= 2/5=0.400$

© 2016 Global Journals Inc. (US)

we obtain the following similarity matrix:

	A	B	C	D	E	F
A	0.000					
B	0.500	0.000				
C	0.500	0.500	0.000			
D	0.167	0.400	0.667	0.000		
E	0.167	0.400	0.667	1.000	0.0000	
F	0.200	0.500	0.500	0.400	0.400	0.000

Figure 5 : Jaccard Similarity Coefficient matrix for the six data vectors in the data matrix shown in Table /1

Jaccard Similarity Coefficient equates similarity with the three types of matches (a, b, c) only, excluding the coefficient of match 'd'. It, however, indicates maximum similarity when the two data vectors have identical values, in which case b=c=0 and $S_{AB=}$1.0. This coefficient also indicates maximum dissimilarity when there are no 1-1 matches, in which case a=0 and $S_{AB=}$0.0. The basic idea of similarity coefficient is to give relative similarity between data vectors. Two data vectors are similar, relative to the cluster membership, if their profiles across variables are "close" or they share "many" characteristics in common, relative to those which other pairs share in common. For the Jaccard similarity coefficient matrix, we obtain the following hierarchical tree:

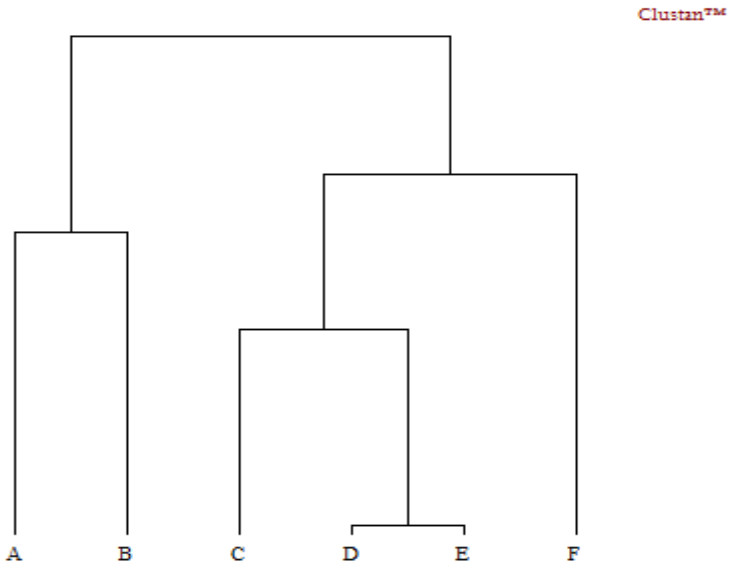


Figure 6 : AHCA using Jaccard Similarity Coefficient

It can be seen that the data vectors (D) and (E) are mathematically most similar (or closest) to each other since they have identical matching coefficients (b and c = 0) and the similarity coefficient between them has the value of 1.0. Data vectors (A) and (B) are also similar since they share similar coefficient of matches. It can also be seen that the data vector (F) is very different from the others. (Note that because only simple data matrix have been used there are only two data vectors representing the two most similar cases that are closer to each other than any other pair in the data matrix).

The similarity coefficients depend on the selected agglomerative clustering method for constructing the hierarchical tree and thus may differ for different methods or different similarity coefficients. Look at the following dendrograms generated by different hierarchical clustering methods using the Jaccard similarity coefficient:

© 2016 Global Journals Inc. (US)

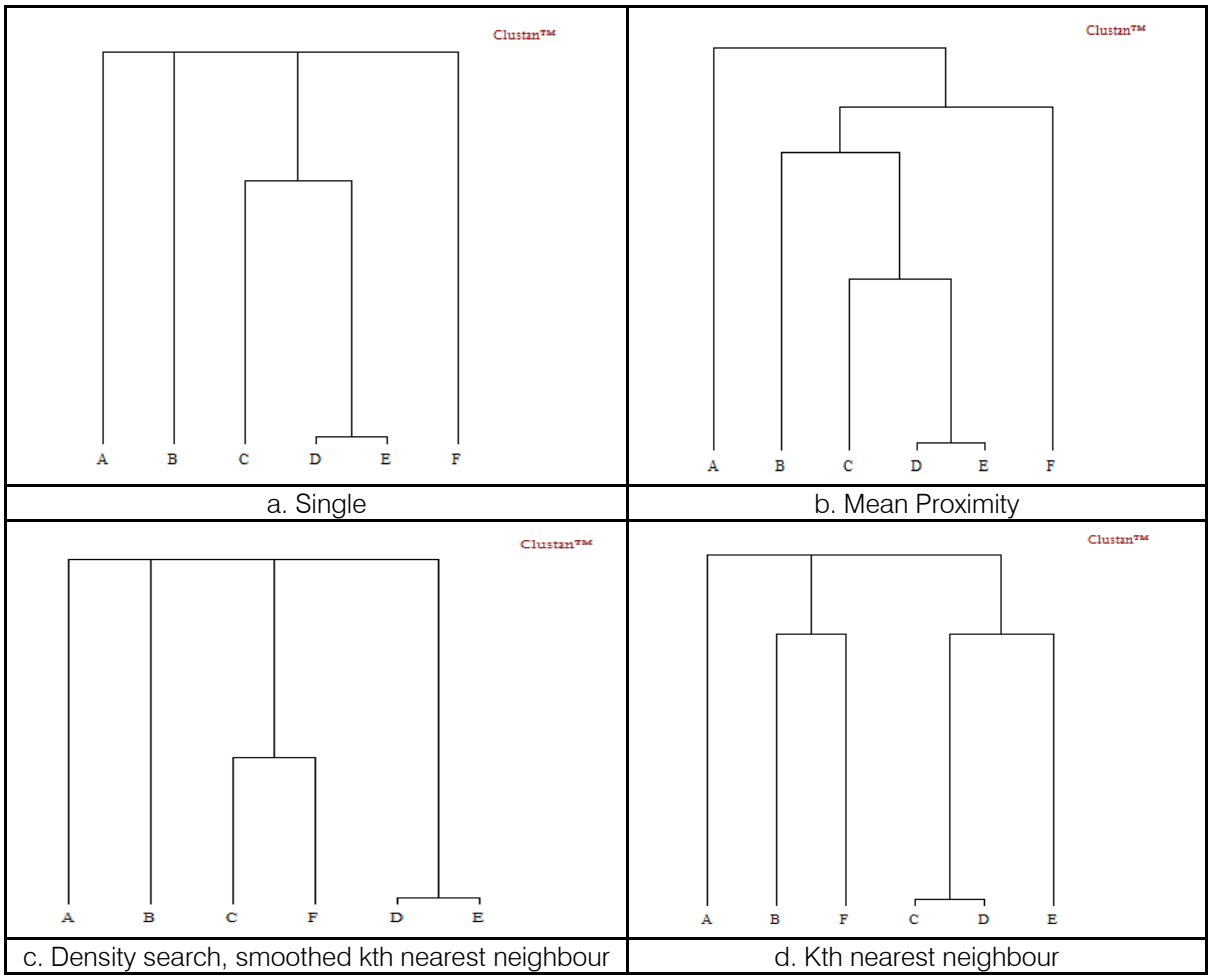


Figure 7 : AHCA four agglomerative hierarchical methods using Jaccard Similarity Coefficient applied to the matrix in Figure 5

More is said about all of these methods in due course; the important thing to realize at this stage is that Jaccard Similarity Coefficient was tried with Ward, Median, Centroid, and Sum of Squares, but the application showed that these methods are not defined for similarity coefficients. To work on it, however, similarity coefficient would have to be converted to dissimilarity by subtracting every value from the maximum similarity by using one of the standard conversion methods:

$$dissim_{ij} = \sqrt{sim_{ii} + sim_{jj} - 2sim_{ij}}$$

(Note when you subtract from the maximum we invert the scale so that previously small values are large. Another way to invert the scale is to multiply the similarity values by minus one, creating dissimilarity values).

The possible similarity coefficients of pairwise similarity are many, and these, together with their equations and properties, are available in, for example, [2], [3], and [10].

Table 3 : various similarity coefficients

Similarity Coefficient	Equation
Matching coefficient	$Sij=(a+d)/(a+b+c+d)$
Jaccard coefficient (Jaccard 1908)	$Sij=a/(a+b+c)$
Rogers and Tanimoto (1960)	$Sij=(a+d)/[a+2(b+c)+d]$
Sneath and Sokal (1973)	$Sij=a/[a+2(b+c)]$
Gower and Legendre (1986 A)	$Sij=(a+d)/[a+1/2(b+c)+d]$
Gower and Legendre (1986 B)	$Sij=a/[a+1/2(b+c)]$
Yule coefficient	$Sij=ad-bc/ad+bc$

© 2016 Global Journals Inc. (US)

Hamann coefficient	$Sij=(a+d)-(b+c)/(a+d)+(b+c)$
Sorenson coefficient	$Sij=2a/2a+b+c$
Rusell and Rao coefficient	$Sij=a/a+b+c+d$

However, when all the selected variables are numerical (continuous or discrete), distance between all pairs of data vectors is commonly computed by using a suitable distance coefficient. A distance coefficient is a measure which defines a distance between vectors of a set of data and it is typically termed metric space if it achieves the metric (triangular) inequality. Ideally, every distance measure should be a metric if the following conditions are satisfied:

$d(x,y) \geq 0$: this condition defines a positive-definite function, saying that distance can't be negative.

$d(x,y)=0$ if $x=y$: this condition says, as above, that distances are always positive except where the data vectors are identical in which case the distance is necessarily 0.

$d(x,y)=d(y,x)$: this condition says that the distance from x to y is the same as the distance from y to x, i.e. the distance is symmetric.

$d(x,z) \leq d(x,y)+d(y,z)$: this condition is called the triangle inequality which says that for any triangle, the sum of the lengths of any two sides must be greater than or equal to the length of the remaining side. The triangle inequality can only be an equality if the remaining side lies exactly on the line connecting the two sides.

In mathematics, a metric space is a set for which distances between all data vectors in the set are defined. These distances, taken together, are called a metric on the set. A distance coefficient is said to have the Euclidean property or to be Euclidean if it always produces distance matrices that are fully embedded in a Euclidean space (i.e. points in space). If a distance matrix is Euclidean then it is also metric but the converse does not follow. Non-Euclidean distances are of different kinds: some still satisfy the metric inequality but have no Euclidean representation (e.g. City block distance), while others are not (e.g. Bray-Curtis distance). The application of these distance measures in agglomerative clustering still makes very good sense as a distance measure between different objects. Discussions on non-Euclidean distances and their applications can be found in, e.g. [11] and [12].

However, choices for some of these distance coefficients are given in the following table that summarizes their equations and properties:

Table 4 : Distance coefficients

Distance coefficient	Description
Squared Euclidean Distance	This measures the distance d between two data vectors i and j, and is expressed as: $$d_{ij}^2 = S_k \frac{W_{ijk}\ (x_{ik}-x_{jk})^2}{S_k\ w_{ijk}}$$ where: X_{ik} is the value of variable k in data vector i, and W_{ijk} is a weight of 1 or 0 depending upon whether or not the comparison is valid for the kth; if differential variable weights are specified. It is the weight of the kth variable, or 0 if the comparison is not valid.
Euclidean Distance	This measures the distance d_{ij} which is obtained by taking the Square root of Squared Euclidean Distance d_{ij}^2 as calculated above.
Euclidean Sum of Squares	The Euclidean Sum of Squares (ESS) EP for cluster P is expressed by: $$E_p = S_{iep} c_i\ S_j \frac{W_j (x_{ij}-mp_j)^2}{S_j\ w_j}$$ Where: X_{ij} is the value of variable j in data vector i within cluster P Ci is an optional differential weight for data vector i W_j is an optional differential weight for variable j m_{pj} is the mean of variable j for cluster P The total ESS for all clusters P is thus $E=S_pE_p$ and the increase in the Euclidean Sum of Squares $I_p E_q$ at the union of two clusters p and q is: IpEq = Ep Eq-Ep-Eq
City Block Distance	City Block Distance, or the Manhatten metric distance, is the Sum of the distances on each variable and is expressed as:

© 2016 Global Journals Inc. (US)

	$d_{ij} = S_K \frac{W_{ijk} \left\| x_{ik} - x_{jk} \right\|}{S_k \; w_{ijk}}$
Product-Moment Correlation	Pearson's correlation coefficient gives the correlation coefficient distance between vectors A and B, and is expressed as: $S_{i,j} = \sum_{K-1}^{N} (C_{k,i} - \overline{C}_i)(C_{k,j} - \overline{C}_j) \Big/ \sqrt{\sum_{K-1}^{N} (C_{k,i} - \overline{C}_i)^2 \sum_{K-1}^{N} (C_{k,j} - \overline{C}_j)^2}$

These distances are closely related, and if all the variables are measured on the same scale or have been transformed or standardized, there is no particular reason to prefer one over another. But if all the variables are measured on the different scale or if the data comprise different variables, then it is important to select the most appropriate proximity coefficient prior to clustering. Detailed discussion on distances in vector space can be found in, e.g., [13] and [14].

c) Vector space

The central concept in agglomerative hierarchical clustering is data vectors in n-dimensional vector space. To understand how hierarchical clustering works, it is necessary to have a firm grasp of this concept. For the present purpose, the distance measure that is most commonly used, most straightforward to apply, and practically simple to understand, will be sufficient. This is the Euclidean distance, or straight-line distance, and almost everyone is familiar with, i.e. can be measured with a ruler.

A Euclidean vector space is a geometrical interpretation of a vector in which the dimensionality *n* of the vector defines an *n*-dimensional space, the sequence of numerical values comprising the vector specifies coordinates in the space, and the vector itself is a point at the specified Cartesian coordinates [1], [15], [16], and [17]. For example, a vector **v** = (2, 4) defines a two-dimensional space and its two components are coordinates in that space; a vector **v** = (2,4,6) defines a 3-dimensional space, and its values in the specified coordinate system place it at the corresponding position in the space; and so on to any dimensionality. This is shown graphically in Figure/8:

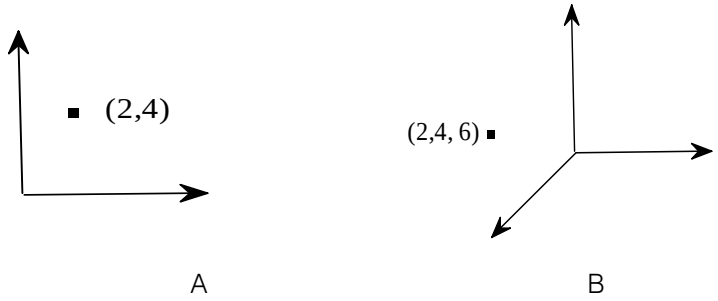


Figure 8 : 2 and 3-dimensional vector spaces

Any number *m* of vectors can exist in an *n*-dimensional vector space, where *m* corresponds to the number of rows in any given matrix M, and *n* corresponds to the number of columns.

d) Distance in vector space

In what follows, the generic term "proximity" is used to refer to the distance relations between and among pairs of vectors. This may be understood in the following ways.

To speak of a vector as a straight line, we see that if we draw a straight line from the origin (0,0) to the position of any point in the space of the axes (X,Y), the distance between the origin to that point is known as the length of a vector and can be measured as in Figure/9.

© 2016 Global Journals Inc. (US)

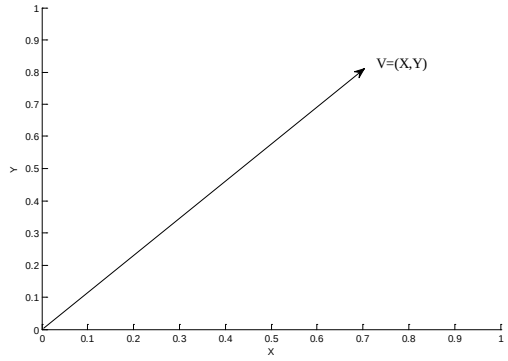


Figure 9 : A Vector in space

If we draw two straight lines from the origin (0,0) to the position of point A and B then we know that there are two vectors in the space and their lengths can be measured and compared. Two straight lines (vectors) are called equivalent (equal) if they have the same length, and unequal if they have different length. Thus the figure/10 shows that the length of vector A is greater than the length of B.

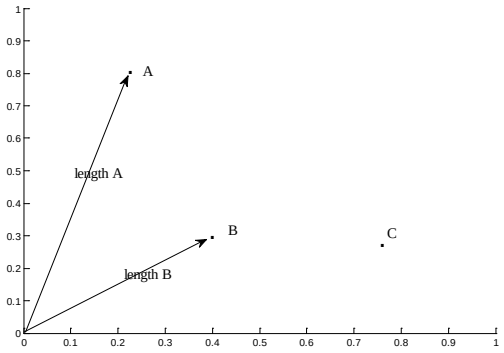


Figure 10 : Vector length

Because each vector is understood as a straight line determined by 2 points in the coordinate system, we may find the position of any vector if its coordinates are known (i.e. the position of vectors with reference to those two lines is known when we know their distances from the axes). Thus, in the figure/10 the position of the vector A is (0.2, 0.8) and vector (B) is (0.4,0.3).

Based on geometrical notions, we may state that the basic elements of vector space are length and angle. These can be used to determine the distance relations between and among vectors, and thus their cluster structure. To illustrate this, when two straight lines (or vectors) meet at a point in a space, there is an angle θ between them, as shown in the Figure/11 below.

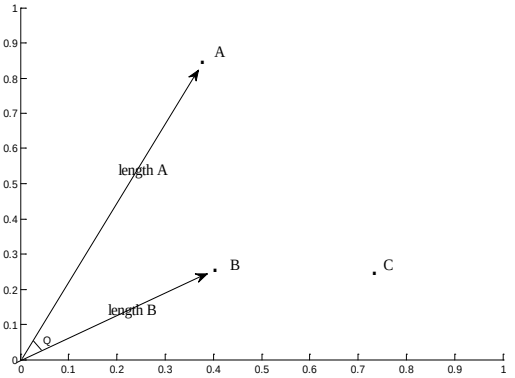


Figure 11 : The angle between vectors

After the length and angle are identified, the distance between two vectors can be measured and relative distances between pairs of vectors compared, so that distance (AC) in figure/12 is greater than distance (AB); this is the basis for several types of clustering method.

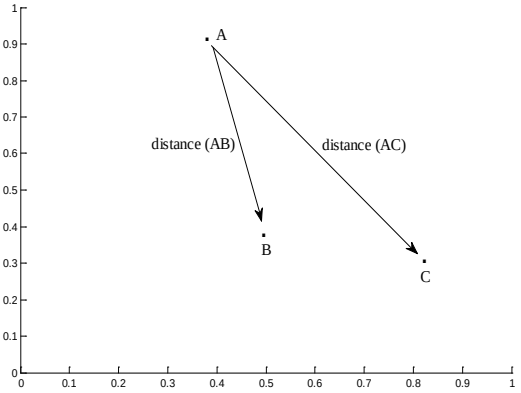


Figure 12 : Vector distances

The distance between any two vectors in a space is determined by the size of the angle between the straight lines meeting at the main point or origin of the space's coordinate system, and on the lengths of those lines. Suppose A and B to be any two vectors having identical lengths and separated by an angle θ (figure/13):

© 2016 Global Journals Inc. (US)

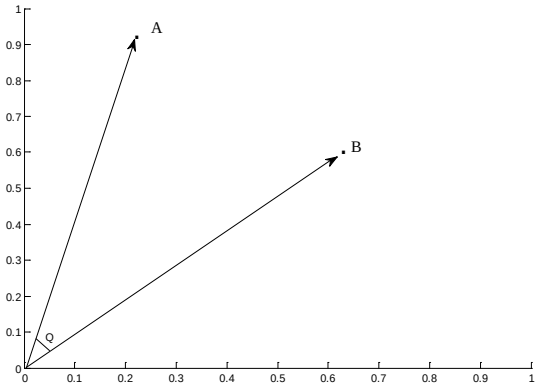


Figure 13

If the angle is fixed and the lengths of the vectors are not the same, then the distance between the two vectors A and B increases (figures/14a and 14b).

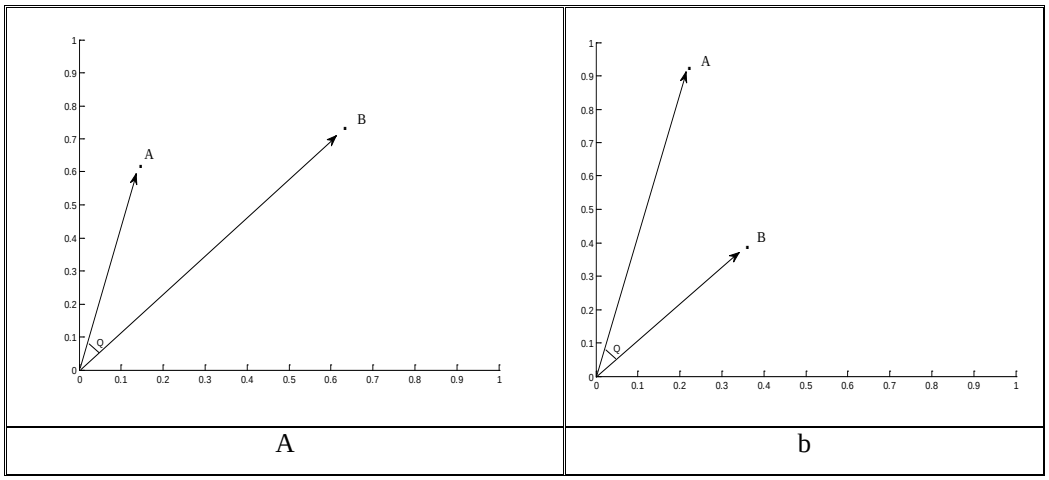


Figure 14

If the lengths of the vectors are the same but the degree of the angle is increased, the distance between the vectors increases (figure/15a), and if the degree of the angel is decreased, the distance is also decreased (figure/15b).

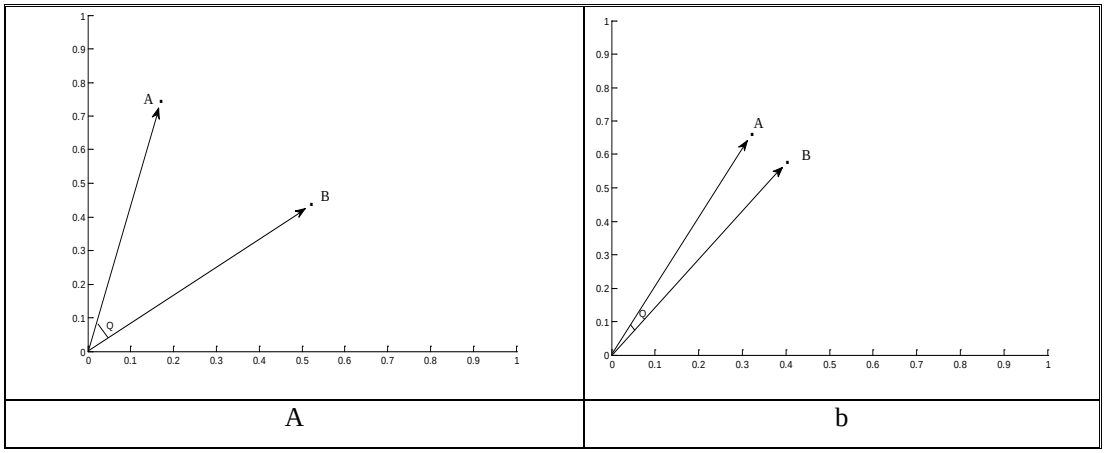


Figure 15

© 2016 Global Journals Inc. (US)

e) Distance in vector space

Most agglomerative hierarchical clustering methods however rely on the concept of distance among data vectors in n-dimensional space (data is represented in the form vectors of real numbers). Data vectors are grouped into similar or dissimilar clusters based on the information found in them: data vectors are considered similar if they are closer together and dissimilar if they are further apart in n-dimensional space. An intuition for how the measure of the distance between vectors in a vector space is best gained by working through a simple numerical example. Very often we use the equation for the Euclidean distance to quantify the distance in vector space. Consider the following triangle:

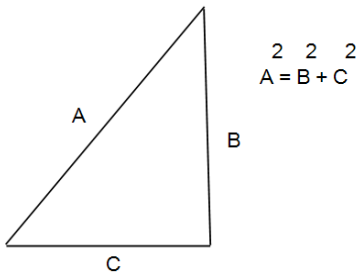


Figure 16 : Intuitive example

Here, the distance between the two points at the vertices of the triangle is:

$$\mathbf{length(A)} = \sqrt{(\boldsymbol{length(B)^2}} + \sqrt{(\boldsymbol{length\ (C)^2}}$$

The origin of this equation is in the Pythagorean Theorem. Pythagoras' theorem says that if we square the two shorter sides in a right-angled triangle and add them together, we get the same as when you square the longest side (the hypotenuse). In the triangle in Figure/16, (B) and (C) are the two shorter sides and (A) is the hypotenuse, so if we square (B) and (C) and add them together B^2+C^2 we get the same as if we square A (A^2). Therefore, $B^2 + C^2 = A^2$. Consider two points in 2- dimensional space:

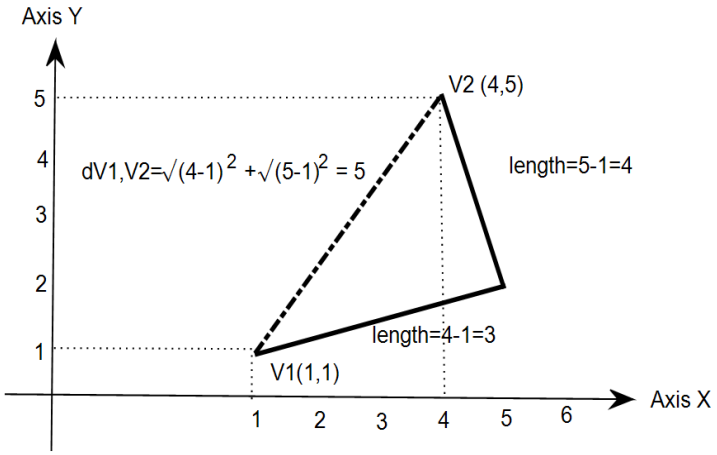


Figure 17 : Pythagoras' theorem applied to distances in two-dimensional space

The horizontal line (i.e. distance) goes from V1 at (1, 1) to V2 at (4,5), so it is obvious that its length $|X1-X2|$ is (4-1)=3 units. The vertical line or distance goes from V2 at (4,5) to (1,1), so again its length $|Y1-Y2|$ is obvious = 4 units. With this in mind, we get a right-angled triangle with lengths 3 and 4. By the Pythagorean theorem, the square of the hypotenuse is $(\text{hypotenuse})^2 = 3^2+4^2 = 25$, which gives the length of the hypotenuse as 25, same as the distance between the two vectors V1 and V2 according to the distance equation above. Thus the Euclidean distance between them is $dV1,V2 = \sqrt{(4-1)^2} + \sqrt{(5-1)^2} = 5$.

Various other distance measures are also possible as discussed above, but they needn't concern us here. Euclidean distance is the simplest and most widely used of the various distance measures. Euclidean distance is also best provided for in software implementations, and so is used here.

However, this quantification applies to any dimensionality n. That is, Euclidean distance applying Pythagoras' theorem can also be generalized or extended to measure the distance between any number of data vectors in any number of dimensions.

$$\boldsymbol{di,j} = \sqrt{(\boldsymbol{i1-j1})^2} + \sqrt{(\boldsymbol{i2-j2})^2} + \sqrt{(\boldsymbol{i3-j3})^2}$$

© 2016 Global Journals Inc. (US)

Look at the figure/18which shows 9 data vectors forming four triangles in 3-dimensional space, where each triangle is in its own space.

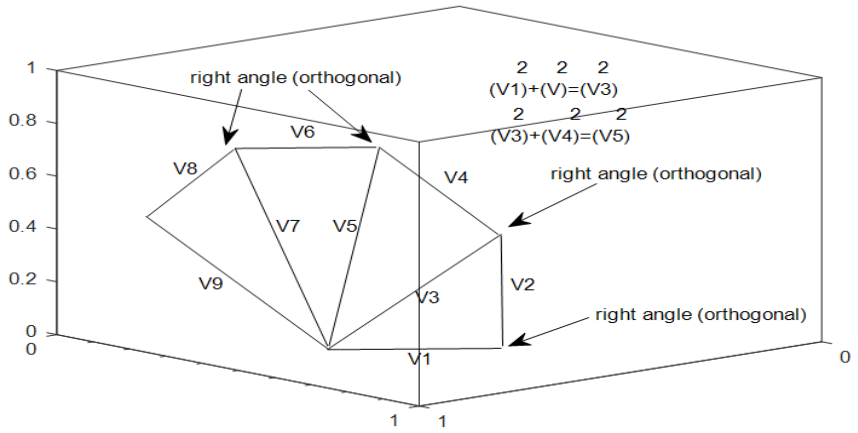


Figure 18 : 4 triangles in 3-dimensional space based on Pythagoras' theorem

More triangles can be found based on the distance measurements among the 9 data vectors but in this figure we limit the calculation to four triangles and the dimensionalities to three.

f) Distance matrix and agglomerative clustering

Because the above quantification of distance in vector space applies to any dimensionality, and not just to the 2 and 3-dimensional spaces that can be visualized, it can be used to define clusters in data of any dimensionality. This is what agglomerative hierarchical clustering does, and it does so in two steps:

i. *Construction of a distance matrix*

When all the distances between all possible pairs of data vectors are measured, they are gathered and entered in a distance matrix which looks like the Table 5:

Table 5 : Distance matrix based on Euclidean distance between 4 data vectors

	V1	V2	V3	V4
V1	0	2.828	3.162	5.99
V2	2.828	0	1.414	3.162
V3	3.162	1.414	0	2
V4	5.099	3.162	2	0

Looking at the distance matrix shows that all of the entries on the main diagonal are zero because the distance from a data vector to itself is zero and that the stored values in the triangle below the diagonal are mirror-images of the stored ones in the triangle above. The distance matrix is an n x n symmetrical, with rows and columns, on either side since the distance between V1 and V2 is identical to the distance between V2 and V1: the distance between any pair of vectors is the same in either direction.

ii. *Construction of a hierarchical tree based on the distance matrix*

Agglomerative hierarchal cluster analysis uses the quantified notion of distance described above, and the distance table more particularly, to find clusters in data. Numerous ways of doing this has been developed, most of them are variations on a theme; for present purposes the theme goes like this.

- For a data set containing m vectors, we start by defining m clusters, one for each vector.
- Using as many steps as necessary, at each step we combine the two clusters with the smallest distance between them into a new, composite (sub) cluster.

To understand this, consider the following data that consists of 14 two-dimensional points shown in Table 6.

Table 6 : a 14 x 8 data matrix

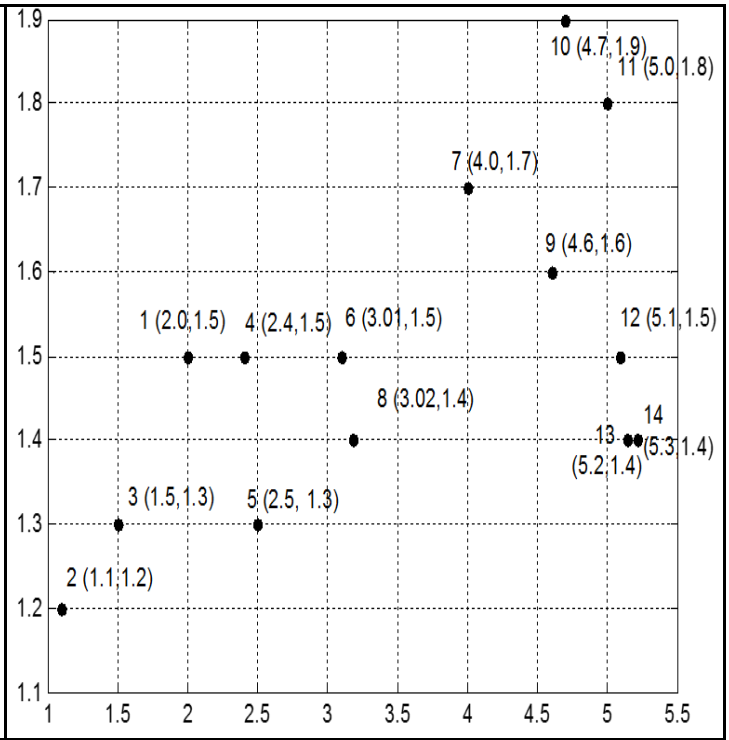
1	4	1.10	1.09	1.79	0.99	1.14	3.25
2	4	1.20	1.08	1.61	0.99	1.15	3.24
3	4	1.19	1.07	1.62	1.15	1.23	3.27
4	4	1.18	1.06	1.61	1.98	1.16	3.22
5	4	1.16	1.04	1.64	0.96	1.17	1.21
6	0.94	0.43	0.38	2.00	0.97	1.06	0.80
7	0.96	0.47	0.43	1.44	0.97	1.10	0.87
8	0.94	0.47	0.43	1.79	0.95	1.10	0.88
9	0.94	0.92	0.84	1.77	0.98	1.14	0.93

© 2016 Global Journals Inc. (US)

10	0.98	0.79	0.76	1.47	0.96	1.12	0.13
11	0.99	0.49	0.47	0.01	0.99	1.13	0.08
12	2.00	3.50	3.49	3.02	0.83	1.13	4.14
13	2.02	3.40	3.72	3.16	0.97	1.19	4.18
14	2.04	3.52	3.52	3.24	0.93	1.12	4.25

The x y coordinates of the points and the plots are shown in Figure/19:

Case	X	Y
1	2.0	1.5
2	1.1	1.2
3	1.5	1.3
4	2.4	1.5
5	2.5	1.3
6	3.0	1.5
7	4.0	1.7
8	3.02	1.4
9	4.6	1.6
10	4.7	1.9
11	5.0	1.8
12	5.1	1.5
13	5.2	1.4
14	5.3	1.4



a B

Figure 19 : The xy coordinates of the 14 data vectors (right) of data matrix in Table/6 (left)

We calculate the Euclidean distance between all pairs of vectors as shown in Figure/17 above and construct the distance matrix for the 14 vectors. Figure/20 below is the one that we looked at above in Table/6 and it is repeated here for clarity of the indicated area between point 3 and 1.

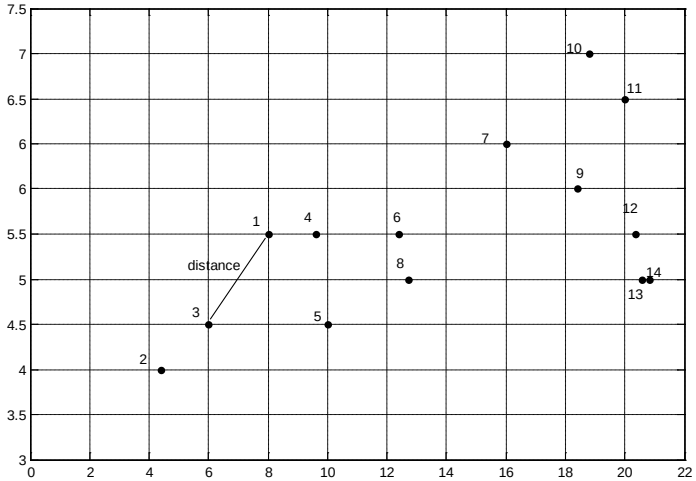


Figure 20 : The plot of the 14 data points

© 2016 Global Journals Inc. (US)

For this data matrix, we abstract the following distance matrix:

Table 7 : A distance matrix for the 14 data vectors

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	**0.000**	0.006	0.010	0.146	0.599	2.339	2.266	2.259	2.120	2.738	3.291	2.550	2.696	2.673
2	0.006	**0.000**	0.005	0.140	0.589	2.366	2.264	2.275	2.123	2.726	3.212	2.560	2.717	2.694
3	0.010	0.005	**0.000**	0.100	0.612	2.389	2.287	2.299	2.146	2.758	3.245	2.575	2.723	2.703
4	0.146	0.140	0.100	**0.000**	0.726	2.490	2.388	2.405	2.250	2.853	3.326	2.778	2.896	2.884
5	0.599	0.589	0.612	0.726	**0.000**	1.520	1.464	1.478	1.356	1.505	1.967	3.712	3.893	3.910
6	2.339	2.366	2.389	2.490	1.520	**0.000**	1.464	1.478	1.356	1.505	1.967	3.712	3.893	3.910
7	2.266	2.264	2.287	2.388	1.464	0.046	**0.000**	0.018	0.069	0.109	0.382	4.691	4.922	4.955
8	2.259	2.275	2.299	2.405	1.478	0.008	0.018	**0.000**	0.054	0.125	0.545	4.546	4.764	4.789
9	2.120	2.123	2.146	2.250	1.356	0.075	0.069	0.054	**0.000**	0.108	0.592	3.813	4.015	4.048
10	2.738	2.726	2.758	2.853	1.505	0.144	0.109	0.125	0.108	**0.000**	0.330	4.905	5.131	5.186
11	3.291	3.212	3.245	3.326	1.967	0.643	0.382	0.545	0.592	0.330	**0.000**	6.396	6.690	6.773
12	2.550	2.560	2.575	2.778	3.712	4.634	4.691	4.546	3.813	4.905	6.396	**0.000**	0.015	0.011
13	2.696	2.717	2.723	2.896	3.893	4.847	4.922	4.764	4.015	5.131	6.690	0.015	**0.000**	0.010
14	2.673	2.694	2.703	2.884	3.910	4.866	4.955	4.789	4.048	5.186	6.773	0.011	0.010	**0.000**

In what follows a 6 x 6 subset of the original 14 x14 distance matrix constructed in Table/7will be used. This makes it possible to show the whole process of constructing a hierarchical tree step by step rather than just a fragment, thereby baking the discussion clearer. The procedure is based on the principal that a set of data vectors has a cluster structure if it can be divided into two or more groups in which the members of any given group are close to one another in the data space, and far from members of other cluster in the space. At each step in tree construction, therefore, one looks for the clusters that are closest to one another and amalgamates them into a super ordinate cluster, and this continues until all the data vectors have been assigned to one of the clusters. The following discussion will demonstrate this.

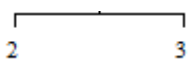
Initially, each row vector of the data matrix is taken to be a cluster on its own; i.e., clusters here and henceforth are shown in brackets. The distance matrix is now searched to find the smallest distance between these data vectors. This is the distance between vector 3 and vector 2 in Table 8: 0.005, shown shaded in Figure/21a. These are combined into a first agglomerated cluster (2, 3) by drawing the tree, as below, and then transforming the distance matrix to incorporate the first cluster.

	1	2	3	4	5	6
1	**0.000**					
2	0.006	**0.000**				
3	0.010	0.005	**0.000**			
4	0.146	0.140	0.100	**0.000**		
5	0.599	0.589	0.612	0.726	**0.000**	
6	2.339	2.366	2.389	2.490	1.520	**0.000**

a. Initial distance matrix reproduced from Table (8) with smallest distance highlighted

Cluster 1	Cluster 2	Agglomerated distance
(2)	(3)	0.005

b. Table of agglomeration



c. Graphical representation of (c)

Figure 21

Transformation of the distance matrix takes a bit of understanding, so it is described in detail.

- The table in figure 21 a is transformed into the one in figure 21b.
- Row vectors and column vectors are removed from the distance matrix and replaced them with a single blank row and column to represent the (2,3) cluster; 0 is inserted as the distance from (2,3) to itself.
- The minimum distances from (2,3) to the remaining data vectors (1), (4), (5), and (6) are inserted into the blank cells of the (2,3) row and column. Confused?

© 2016 Global Journals Inc. (US)

In the original distance matrix, the distance between (2) and (1) is 0.006 and between (3) and (1) is 0.010, shown shaded in figure/22a below. The minimum distance here is 0.006, and is inserted into the relevant cell representing the minimum distance between (2,3) and (1). The distance between (2) and (4) in the original distance matrix is 0.140 and between (3) and (4) it is 0.100. The minimum distance here is 0.100 and it is inserted into the relevant cell representing the distance between (2,3) and (4). The distance between (2) and (5) in the original distance matrix is 0.589 and between (3) and (5) it is 0.612.The minimum distance here is 0.589 and it is inserted into the relevant cell representing the distance between (2,3) and (5). The distance between (2) and (6) in the original distance matrix is 2.366 and between (3) and (6) it is 2.389.The minimum distance here is 2.366 and it is inserted into the relevant cell representing the distance between (2,3) and (6). Emendation of the distance table is now complete, and the resulting table is the basis for the next step in the construction of tree. Now the distance table is searched to find the smallest distance between vectors. This is the distance between vectors (2,3) and (1): 0.006. Vectors (2, 3) and (1) are now combined into a new subordinate cluster ((2,3),1) by drawing the tree as below, and then emending the distance table to incorporate the new cluster.

	1	2	3	4	5	6
1	0.000					
2	0.006	0.000				
3	0.010	0.005	0.000			
4	0.146	0.140	0.100	0.000		
5	0.599	0.589	0.612	0.726	0.000	
6	2.339	2.366	2.389	2.490	1.520	0.000

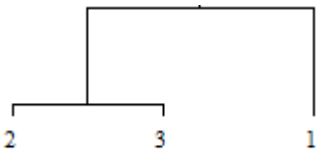
a. Distance matrix from Table (8)

	1	(2,3)	4	5	6
1	0.00				
(2,3)	0.006	0.00			
4	0.146	0.100	0.000		
5	0.599	0.589	0.726	0.000	
6	2.339	2.366	2.490	1.520	0.000

b. Transformed version of a.

Cluster 1	Cluster 2	Agglomerated distance
(2)	(3)	0.005
((2),(3))	(1)	0.006

c. Table of agglomeration



d. Graphical representation of (c)

Figure 22

We must note that the distance matrix has shrunk by one row and column. In any process of agglomerating clusters, this shrinkage will continue as we proceed.

Emendation of the distance table proceeds as step (1) explained above by removing the rows and columns and replacing them with single blank row and column to represent the new ((2,3)1) sub-cluster. Then the minimum distance from ((2,3),1) to the remaining data vectors (4), (5), and (6) is inserted into the blank cells. From Figure/ 22, the distance between (2,3) and (1) is 0.006 and between (4) and (1) is 0.146; the minimum distance is 0.006, and it is inserted into the relevant cell. The distance (2,3) and (5) 0.589 and between (1) and (5) is 0.599; the minimum distance here is 0.589, and it is inserted into the relevant cell. The distance between (2,3) and (6) is 2.366 and between (1) and (6) is 2.339; the minimum here is 2.339, and it is inserted into the relevant cell.

© 2016 Global Journals Inc. (US)

	1	2	3	4	5	6
1	0.000					
2	0.006	0.000				
3	0.010	0.005	0.000			
4	0.146	0.140	0.100	0.000		
5	0.599	0.589	0.612	0.726	0.000	
6	2.339	2.366	2.389	2.490	1.520	0.000

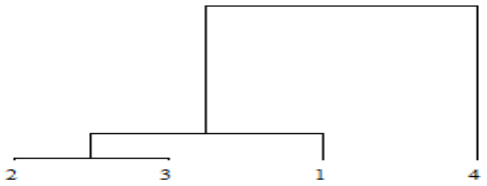
a. Distance matrix from Table (8)

	((2,3),1)	4	5	6
((2,3),1)	0.000			
4	0.006	0.000		
5	0.589	0.726	0.000	0.000
6	2.339	2.490	1.520	0.000

b. Transformed version of a.

Cluster 1	Cluster 2	Agglomerating distance
(2)	(3)	0.005
((2),(3))	(1)	0.006
(((2),(3)),(1))	(4)	0.006

c. Table of agglomeration



d. Graphical representation of (c)

Figure 23

The distance table is searched to find the smallest distance between vectors. This is the distance between vectors ((2,3),1) and (4): 0.006. Clusters ((2,3),1) and (4) are now agglomerated into a subordinate cluster (((2,3),1),4) as shown in the tree above, and then emending the distance matrix to incorporate the new cluster. Emendation of the distance matrix proceeds as in step 1 and 2. The rows and columns (2,3) and (4) are removed from the table and replaced them with a single blank row and column to represent the new (((2,3,4),1) cluster. The next step is to insert into the blank cells the ((2,3),1),4) to the remaining clusters (5) and (6). The distance between ((2,3),1) and (5) is 0.589 and between (4) and (5) is 2.726; the minimum is 0.589 and it is inserted into the relevant cell. The distance between ((2,3),1) and (6) is 2.339 and between (4) and (6) is 2.490; the minimum is 2.339 and it is inserted into the relevant cell. Here the smallest distance is 0.589 and thus clusters ((2,3),1),4) and (5) are now agglomerated into a subordinate cluster (((2,3),1),4),5) as shown in the tree below. The distance matrix is emended to incorporate the new cluster. Emendation of the distance table is now complete and the resulting matrix is the basis for the final step.

	1	2	3	4	5	6
1	0.000					
2	0.006	0.000				
3	0.010	0.005	0.000			
4	0.146	0.140	0.100	0.000		
5	0.599	0.589	0.612	0.726	0.000	
6	2.339	2.366	2.389	2.490	1.520	0.000

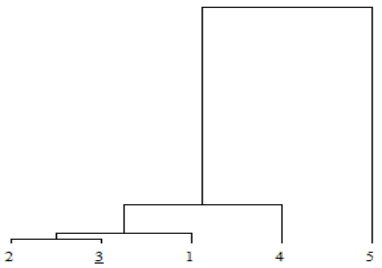
a. Distance matrix from Table (8)

	(((2,3),1),4)	5	6
(((2,3),1),4)	0.000		
5	0.589	0.000	
6	2.339	1.520	0.000

b. Transformed version of a.

Cluster 1	Cluster 2	Agglomerated distance
(2)	(3)	0.005
((2),(3))	(1)	0.006
(((2),(3)),(1))	(4)	0.006
(((2),(3)),(1),(4))	(5)	0.589

c. Table of agglomeration



d. Graphical representation of (c)

Figure 24

© 2016 Global Journals Inc. (US)

The minimum distance from (((2,3),1),4),5) to the remaining vector (6) is inserted into the blank cell of the (((2,3),1),4),5) column. The distance table generated in Figure/21 above is searched to find the smallest distance between vectors. There is only one remaining vector value. Clusters (((2,3)1,4),5) and (6) are now combined into a subordinate cluster (((((2,3),1),4),5),6) by drawing the tree and then emending the distance table to incorporate the new cluster.

	1	2	3	4	5	6
1	0.000					
2	0.006	0.000				
3	0.010	0.005	0.000			
4	0.146	0.140	0.100	0.000		
5	0.599	0.589	0.612	0.726	0.000	
6	2.339	2.366	2.389	2.490	1.520	0.000

a. Distance matrix from Table (8)

	(((2,3),1),4),5)	6
(((2,3),1),4),5)	0.000	
6	2.339	0.000

b. Transformed version of a.

Cluster 1	Cluster 2	Agglomerating distance
(2)	(3)	0.005
((2),(3))	(1)	0.006
(((2),(3)),(1))	(4)	0.006
(((2),(3)),(1),(4))	(5)	0.589
(((2),(3)),(1),(4),(5))	(6)	0.2339

c. Table of agglomeration

d. Graphical representation of (c)

Figure 25

All 6 data vectors have now been incorporated into the cluster tree and tree construction stops.

	(((2,3),1),4),5),6)
(((2,3),1),4),5),6)	0.000

Figure 26

In this example, we only obtained distance measurements and cluster agglomerations for only 6 data vectors from the original 14 x 8 data matrix of Table/7, because the calculation can become extremely long, it is important to emphasize that for a set of 14 data vectors there would be a total of 91 steps including the main diagonal zero-values. This can be given in relationship of the number of possible successive agglomerations: n (n-1)/2 where n is the number of data vectors. However, the steps explained above are repeated on the whole data matrix, and the result is shown in Figure/27:

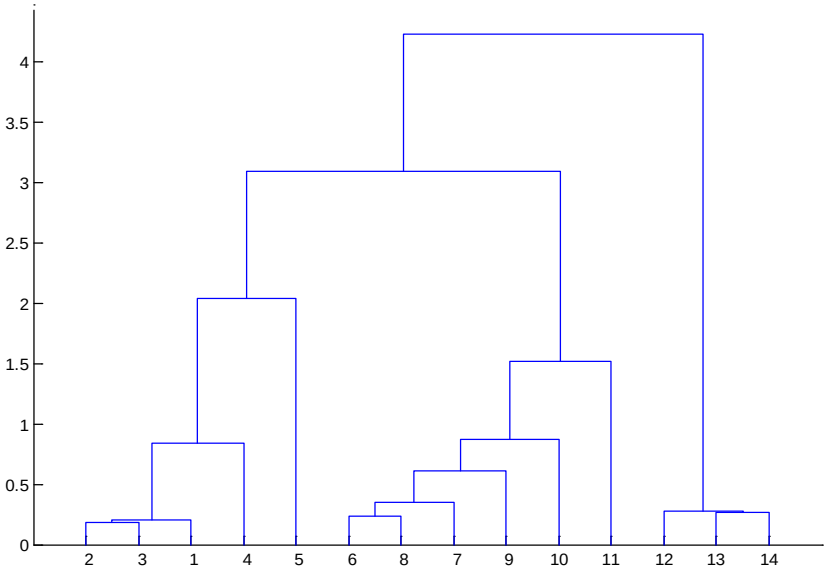


Figure 27 : Agglomerative hierarchical analysis for the 14 data vectors of the data matrix in Table 7

© 2016 Global Journals Inc. (US)

In this figure, the 14 vectors are represented as clusters and agglomerated together on the basis of the relativities of distance among them and the structure presented in a tree-like diagram. In this figure, all the 14 vectors are agglomerated into three main clusters which represent the relativities of distance among them as a dendrogram in figure/ 28a and their corresp.

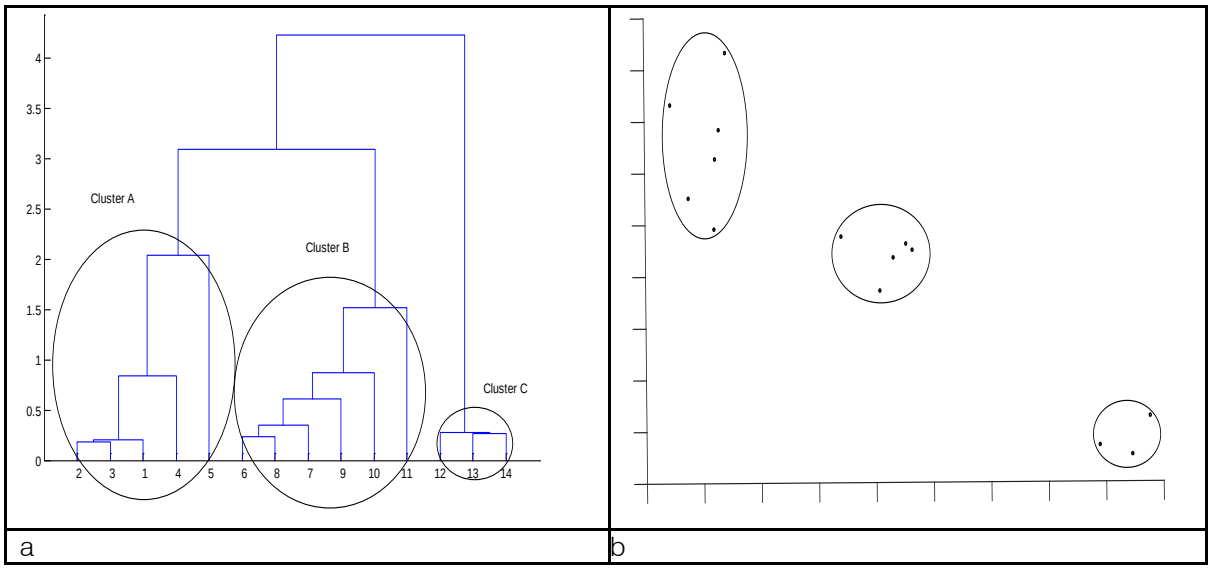


Figure 28 : Three main clusters for the 14 data vectors of the data matrix in Table 7

Given that the hierarchical clustering tree tells us nothing more than what the two-dimensional plot tells us, what is gained? In the current case nothing. The real power of agglomerative hierarchical cluster analysis consists in its independence of vector space dimensionality. Put it another way, direct plotting is limited to two, three, or fewer dimensions but there is no dimensionality limit on agglomerative hierarchical cluster analysis. It can determine relative distances in vector spaces of any clustering and represent those distance relativities as a dendrogram like the one above.

g) Agglomerative Hierarchical Clustering Methods

Many agglomerative clustering methods are treated as variations on a single major approach; they require the data to be in the form of vectors of real numbers and follow the same standard framework: Initially, before clustering has begun, each data vector is treated as a cluster or group, clustering begins by a successive agglomeration of the two closest or nearest pair of clusters (i.e. the two data vectors that are separated by the smallest distance) to form first cluster. The process of agglomerating two data vectors and fusing their characteristics is repeated until only one cluster remains. Extensive range of agglomerative clustering methods exists; though most of them operate in a similar way, their calculation is different. Eleven of these methods are introduced. They are:

- Single linkage (or nearest neighbor) method

In this method, the distance between two clusters A and B is based on the membership (i.e. data vectors) in each cluster that are nearest together (shortest distance).

$$D_s(A,B) = \min\{d(a \in A, b \in B)\}$$

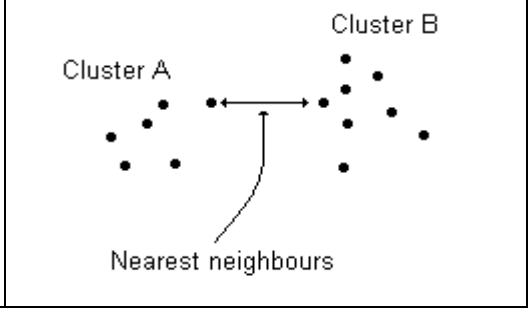


Figure 29 : Single clustering

© 2016 Global Journals Inc. (US)

On this basis, at each step of the clustering process, we combine the two data vectors that have the smallest single linkage distance.

- Complete clustering (furthest neighbor) method

In this method, the distance between two clusters A and B is based on the data vectors in each cluster that are furthest apart or furthest neighbors (longest distance).

$$D_c(A,B) = \max\{d(a \in A, b \in B)\}$$

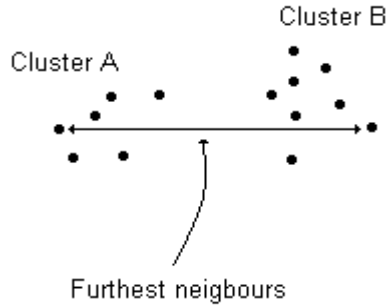


Figure 30 : Complete clustering

On this basis, at each step of the clustering process, we combine the two data vectors that have the smallest complete linkage distance.

- Average clustering method

In this method, also known as the unweighted pair-group using average approach conventionally abbreviated (UPGMA), the distances between all possible data vectors embedded in the two clusters A and B are calculated and summed, and the distance between cluster A and cluster B is the average of that sum.

$$D_{avg}(A,B) = \frac{\sum_{i=1..m, j=1..n} d(a_i \in A, b_j \in B)}{m \times n}$$

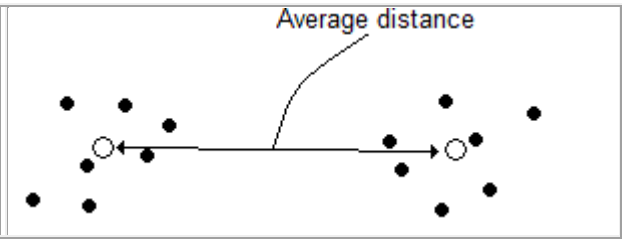


Figure 30 : (Group) Average clustering method

Where $D_{avg}(A,B)$ is the average link distance between A and B, d is the distance between a single pair of data vectors, m is the cardinality of cluster A, and n is the cardinality of cluster B. On this basis, at each step of the clustering process, we combine the two data vectors that have the smallest average linkage distance.

- Weighted Average clustering method

This method has also been referred to as the weighted pair-group using average approach conventionally abbreviated (WPGMA). In this method, when two clusters A and B are agglomerated, the distance D between some other cluster, say, C and the newly formed cluster AB is the simple average of D_{CA} and D_{CB}, thus:

$$(D)\ C, AB = \tfrac{1}{2}\{D_{CA} + D_{CB}\}$$

On this basis, at each step of the clustering process, we combine the two data vectors that have the smallest weighted average linkage distance.

- Ward's method or an increase in sum of squares clustering method

This method involves the concept of sum-of-squares error, abbreviated SSE. Given a set D of *n* values, the SSE of D is the sum of the squared differences between each value in D and the mean of all values in D:

$$SSE_D = \sum_{i=1..n} \left| d_i \in D - \frac{\sum_{j=1..n} d_j \in D}{n} \right|^2$$

Ward's method calculates the distance between clusters A and B as

$$D_{Ward} = SSE(A,B) - \left(SSE(A) + SSE(B)\right)$$

On this basis, at each step of the clustering process, we combine the two data vectors that have the smallest increase in the sum of squares.

- Sum of squares clustering method

The distance between two clusters A and B is calculated as the sum of the squared distances between

© 2016 Global Journals Inc. (US)

the data vectors of clusters A and B and the centroid of the agglomerated cluster. The sum of squares method is only calculated for squared distances. For a given set of n data vectors, this method seeks to minimize the sum of the squared distances between the data vectors and the centers (or means) of the clusters to which they belong. In this respect, it is very similar to Increase in Sum of Squares (Ward's method) above.

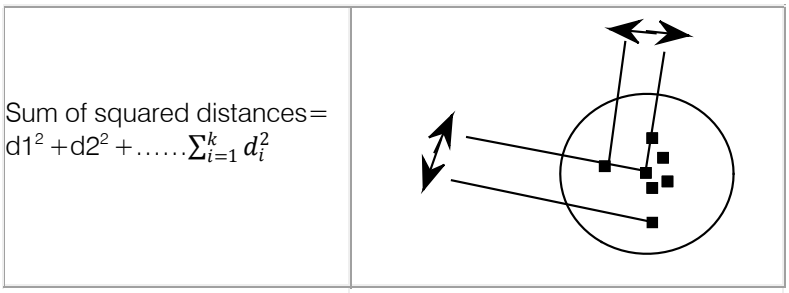


Figure 31 : Sum of Squares clustering

- Centroid clustering method

This method is also known as the unweighted pair-group method using the centroid approach (UPGMC). The Centroid method is only calculated in terms of squared distances. The squared distance between two clusters A and B is calculated as the squared distance between the cluster means, or centroids. The size or weight of a cluster is not relevant, although its spatial distribution is used in the calculation of the centroid. This method should, strictly speaking, only be used with a matrix of squared distances.

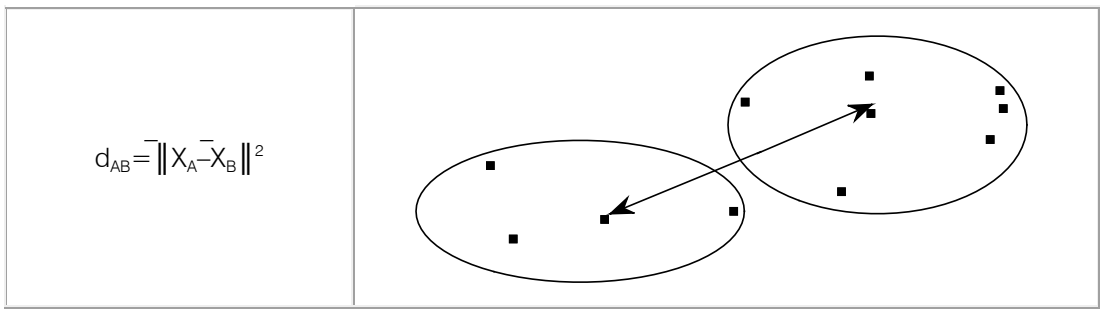


Figure 32 : Centroid clustering

Where X_A and X_B are the mean vectors for the data vectors in A and the data vectors in B respectively. On this basis, at each step of the clustering process, we combine the two data vectors that have the smallest centroid distance.

- Median clustering method

Also known as the weighted pair-group method using centroid approach (WPGMC). The Median method is only calculated in terms of squared distance. In this method, the distance between two clusters A and B is represented by the squared Euclidean distance between the median (mid-point) for the data vectors in cluster A and the median for the data vectors in cluster B. This gives equal weight to clusters of different sizes, unlike the centroid, which is weighted by the number of data vectors in each cluster. However, the two data vectors with the smallest distance between medians are agglomerated at each step.

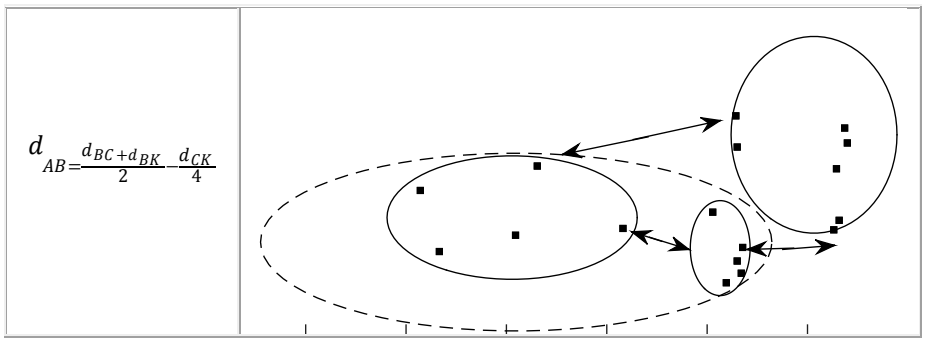


Figure 33 : Median clustering

© 2016 Global Journals Inc. (US)

- Flexible beta clustering method

This method calculates the distance between two data vectors on the basis of β which is a supplied by the user. By allowing β to vary, clustering results with various characteristics can be obtained. However, a value of β = -0.25 gives results similar to Ward's method. A detailed account on the mathematical properties of this method can be found in, e.g., [18] and [19].

- Mean proximity clustering method

This method maximizes the average of the within-cluster distances or minimizes the average of the between-cluster distances, for all cluster comparisons.

- Density search clustering using nearest- neighbor clustering approach

This method falls into a class of clustering methods particularly designed to seek dense patches, regions or areas in the data vectors in a metric space depending on the type of the density estimation to be used. The density nearest neighbor method uses either K^{th} nearest neighbor density estimates or smoothed K^{th} nearest neighbor estimates. The density estimation of the former is based on a fixed number of values and the density estimation for the latter on a large number of values K, where k is the contiguous or the nearest neighbors to the desired point. The distance between two clusters A and B is based on the value specified for K; the estimated value of k controls the amount by which the data are smoothed or unsmoothed to give the density estimate on which the clustering procedure is based: when the value of k is non-increased or small, the density estimation becomes unsmooth or jagged, when the value of k is increased or large, the destiny estimate becomes smoother or less bumpy. To be more precise, the problem is that all K neighbors must be close to the desired point. This may or may not be possible. Theoretically speaking, this is possible when infinite number of data vectors is available, in such a situation the larger the k value the better is calcification (error rate gets closer to the lowest possible error rate for a given classification). Because this is not always possible in practice due to data vectors are finite, K value should be large so that error rate is minimized; too small values of K may lead to noisy decision boundaries and too large may lead to over-smoothed boundaries. That is, K value should be small enough so that only nearby data vectors are included. However, whatever density estimation it may take, this method consists of two main basic steps: initially, a new distance, based on density estimates and adjacencies in the data vectors, is calculated. This step is obtained by: calculating the K^{th} nearest neighbor for the data vectors: given two clusters A and B, the data vectors X_A and X_B are said to be adjacent (the definition of adjacency depends on the method of density estimation), if $D^*(X_A, X_B) \leq D_K(X_A)$ or $D_K(X_B)$. Where D^* is the distance and $D_K(X_A)$ is the kth nearest neighbor distance to data vector (X_B). The distance $D(X_A, X_B)$ between the data vectors X_A and X_B can be obtained as:

$$D(X_A, X_B) = 0, \text{ if } X_A = X_B;$$
$$= \tfrac{1}{2}[D_K(X_A) + D_k(X_B)], \text{ if } D^*(X_A, X_B) \leq D_K(X_A)$$
$$\text{or } D^*(X_A, X_B) \leq D_K(X_B)$$
$$= \infty \text{ otherwise.}$$

Finally, a single linkage clustering method is then applied to the resulted distance D^* to obtain high-density clusters [2], [3], [10], [16], [20], and [21].

A detailed account on the mathematical properties of these methods can be found in, e.g., [5] and [16].

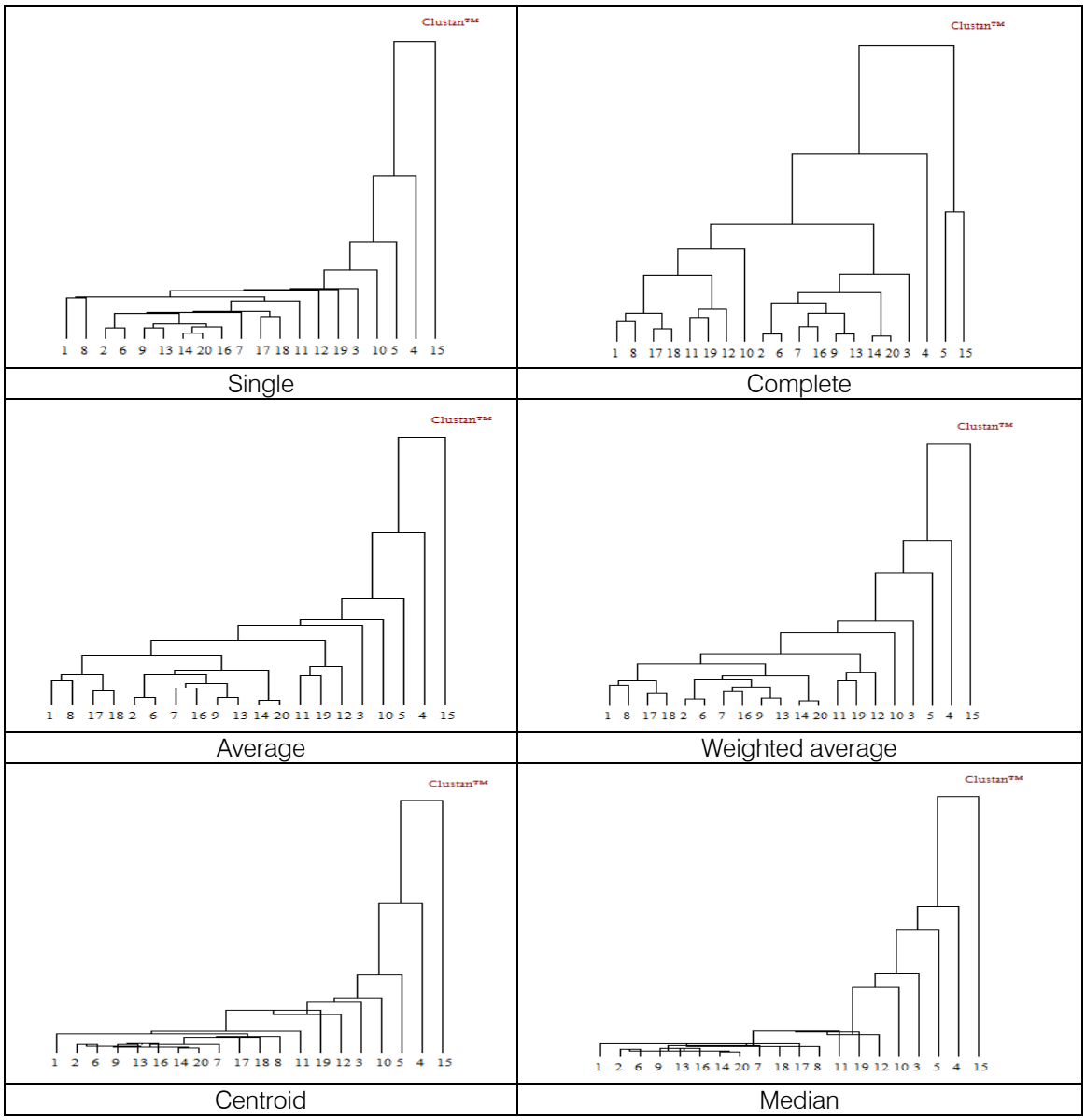
Since the calculation both of the values in the original distance matrix and of the distances between composite clusters are based on linear measurement, agglomerative hierarchical clustering is a collection of linear cluster analysis methods.

Extensive empirical clustering results, however, have shown that, relative to a given data matrix, each agglomerative clustering method has a 'signature' in the sense that the hierarchical tree it produces tend to have specific characteristics [2] and [5].The literature search on the application of hierarchical clustering methods reports, for example, that Single link famously tends to generate 'chained' structures, that is, trees with a strong tendency to either left or right branching but not both. It also reports that this method has satisfactory mathematical properties, which appears to give satisfactory results at identifying longated clusters that have curvy shapes instead of spherical or elliptical shapes, and it is somewhat robust to outliers in the set of data. Complete link tends to generate trees with extensive recursive embedding of left and right branching sub trees; also tends to generate very small compact clusters, which means that they have small diameter (max. distance between data vectors). In other words, group structure, all data vectors in the same cluster, will not be taken into account. On the other hand, this method is somewhat sensitive to outliers, and is suitable for compact but not well-separated clusters. Average linkage is intermediate between single and complete link; it is intermediate between single and complete linkage; it tends to generate small clusters of outliers and to find spherical clusters, i.e. ball-shaped clusters. Being relatively robust, this method can even deal with rather potato-shaped clusters. It is, however, more prone to chaining than Ward's method. Ward's method is like complete link, but in addition tends to find spherical clusters of roughly equal size. As such, some methods are more appropriate than others for data with a given density structure. If, for example, the data manifold has an elongated structure, single link would be best and Ward worst. Alternatively, a manifold with

© 2016 Global Journals Inc. (US)

well-defined spherical areas of vector density would reverse that. Ward's method tends to find spherical clusters of roughly equal size. It is sensitive to outliers. On the other hand, many researchers report satisfactory results with this method (i.e. provides interpretable results). Centroids linkage tends not to chain as much as single linkage. It is nevertheless subject to reversals. Median linkage tends to chain for large set of data and is also subject to reversals. However, they are both fairly robust to outliers. Flexible beta linkage tends to generate 100 % chained clusters if β approaches a value of +1. On the other hand, if β approaches zero and then becomes negative, this method tends to cluster data vectors more intensely. A value of β −0.25 gives results similar to Ward's method. Density nearest-neighbor linkage leads to a very simple approximation of the (most desired) smallest possible error rate for a given classification and data representation. However, it tends to generate different clusters with greater or lesser tendency to chain depending on different values of k. This method tends to overcome the chaining effects if $k = 2\log_2 n$ or several values around this value. On the other hands, this method is prone to produce noisy decisions boundaries. As such some methods are more appropriate than others for data with a given density structure; some methods work better for certain data sets, and other methods work better for other data sets. However, if, for example, the data manifold has an elongated structure, single or nearest neighbor linkage would be best and Ward worst.

As might be expected, different agglomerative clustering methods can and often do give different results for the same dataset. Different clustering structures are obtained when we cluster analysed a data matrix consisting of 20 data vectors applying the 11 methods introduced above.



© 2016 Global Journals Inc. (US)

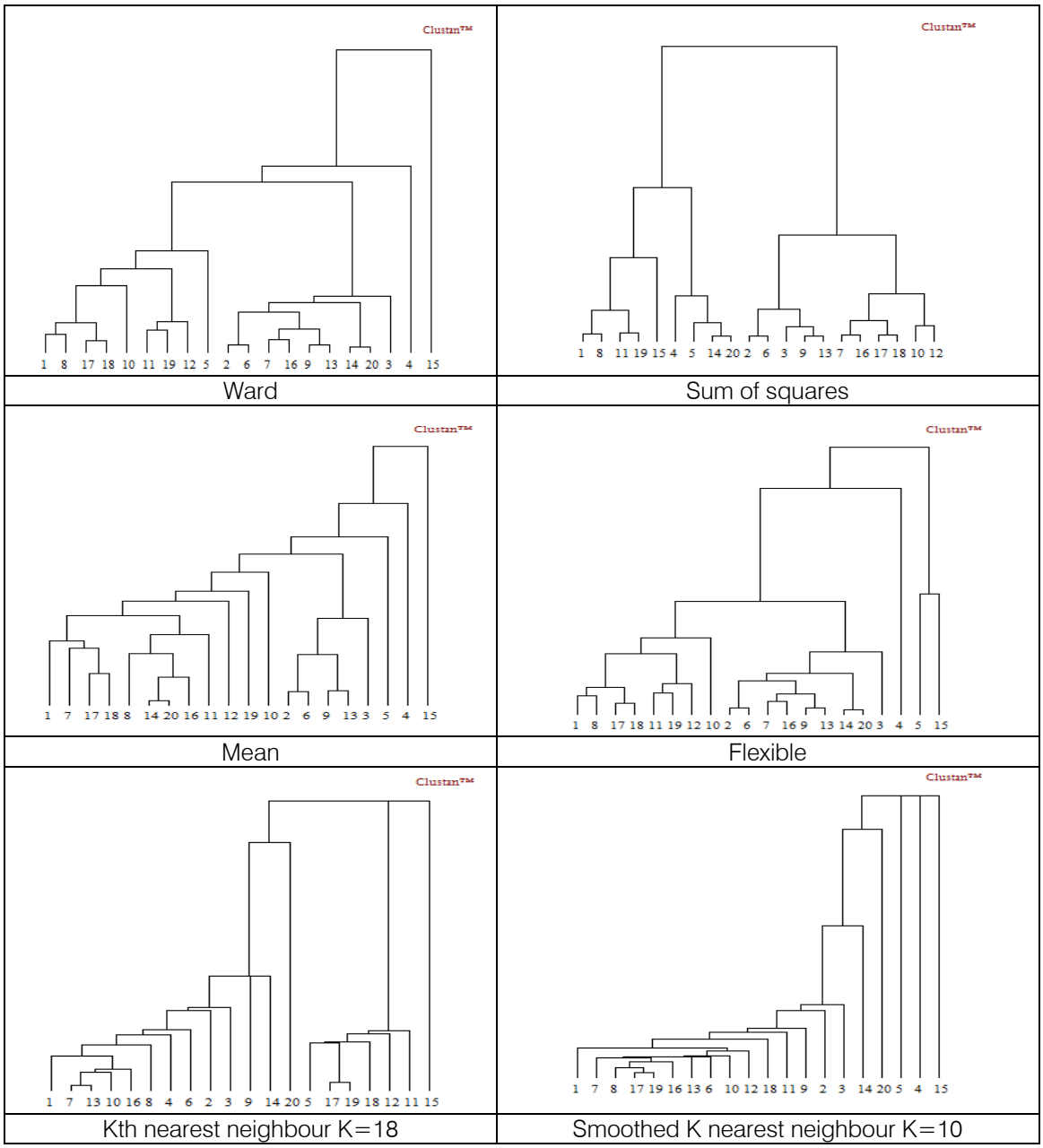
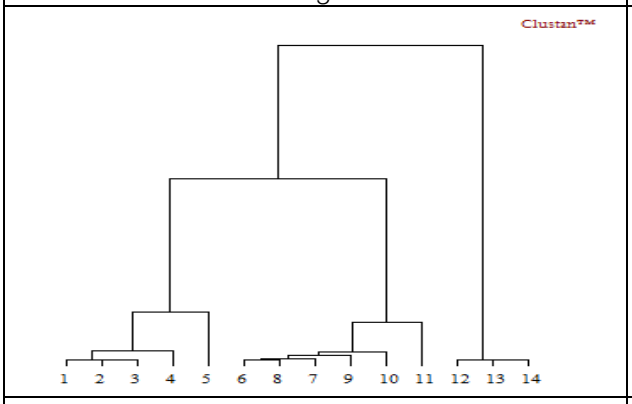
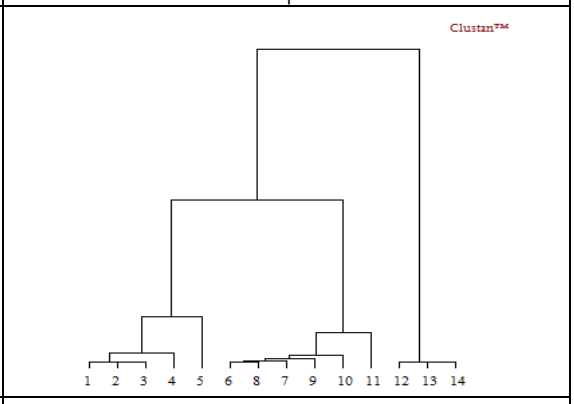
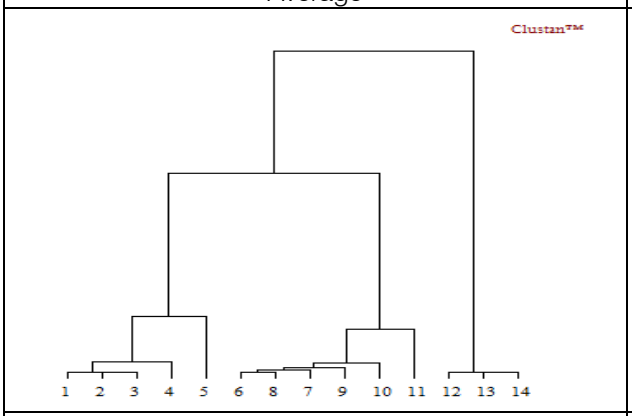
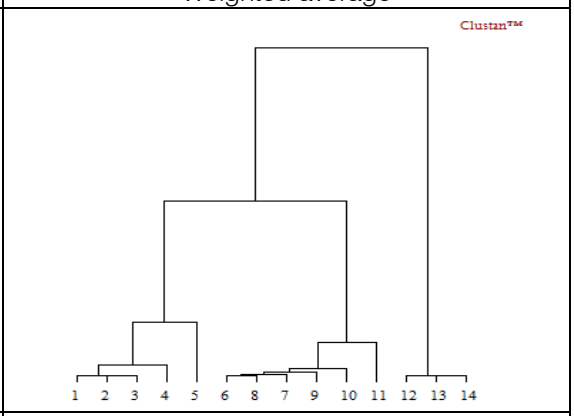
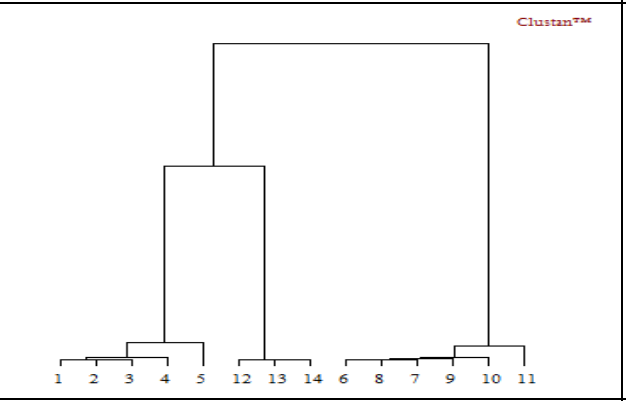
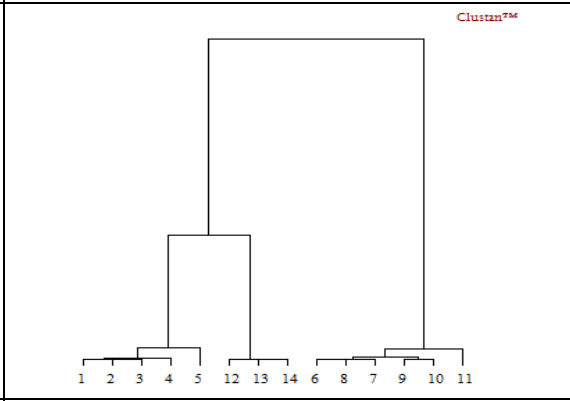
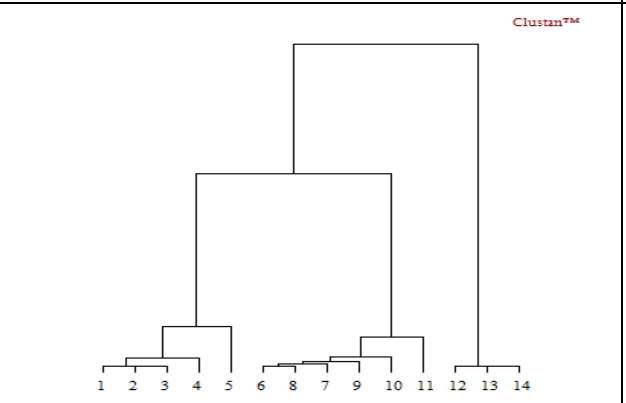
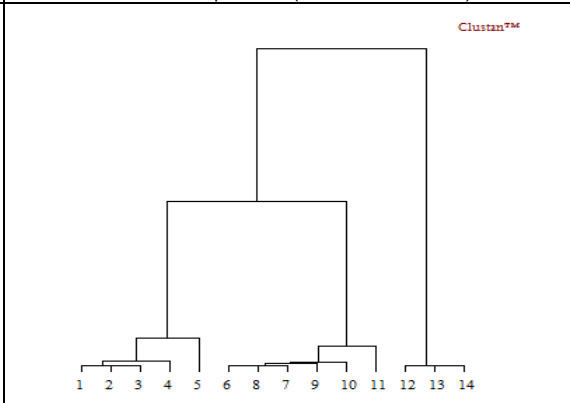


Figure 34 : The application of different hierarchical clustering methods on the same data set using squared Euclidean distance

Which hierarchical analysis is the best? None of these clustering analyses is uniformly the best. In this practice it is advisable to try several methods and then compare the clustering results to form an overall judgment about the final structures of clusters.

Occasionally, however, observed clustering results are very different from those expected. Here is a little example to illustrate this. The following dendrograms generated from the eleven hierarchical clustering methods applied on a small data matrix (i.e. having small measurements).

© 2016 Global Journals Inc. (US)

Single	Complete
	
Average	Weighted average
	
Centroid	Median
	
Ward	Sum of squares (similar to ward)
	
Mean	Flexible

© 2016 Global Journals Inc. (US)

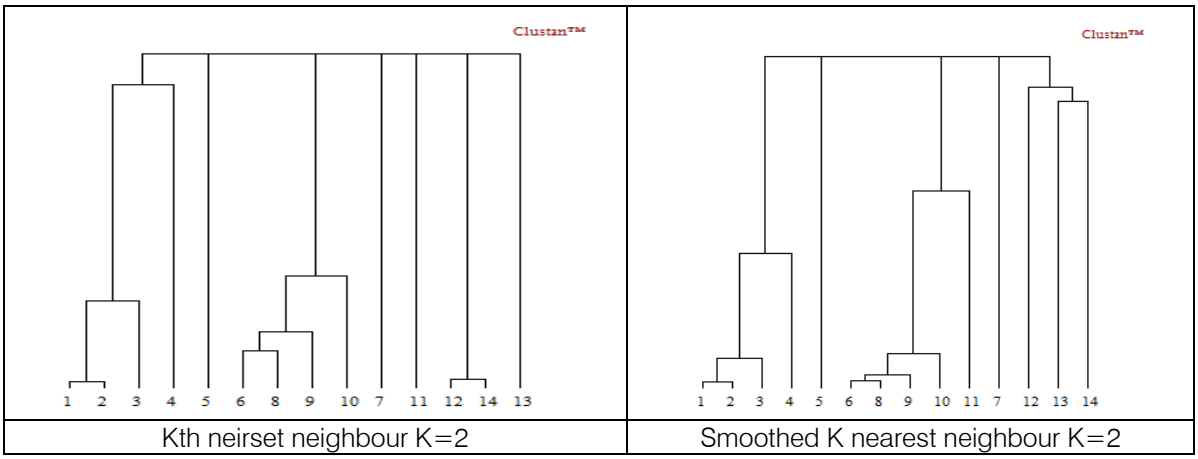


Figure 35 : the application of eleven hierarchical clustering methods on the same (small) data set using squared Euclidean distance

The clustering analyses in this figure show that the application of various agglomerative methods on the same dataset may not always produce quite different results because the clustering results may have generated as a result of highly precise data or a small data matrix size.

III. Conclusion

When using agglomerative hierarchical analysis to form clusters, we need to keep the following in mind:

- Agglomerative hierarchical cluster analysis is a multivariate method for finding structures or groups called clusters in data in relation to a research of interest. The clusters are based on the values of several variable measurements that describe data vectors. The accuracy of agglomerative hierarchical cluster analysis is unquestionable: Data vectors (objects, cases, observations) in a specific cluster share many characteristics, but are very dissimilar to data vectors not belonging to that cluster.
- Prior to analyzing data and applying a clustering method, we need to choose the appropriate proximity coefficient (i.e. measure of distance/ similarity) depending on type of data: interval, counts, binary. Distance is a measure of how far apart two data vectors are, while similarity measures how similar two data vectors are. For data vectors that are similar, distance measures are small and similarity measures are large.
- Proximity coefficients are stored in a proximity matrix. The proximity matrix identifies which cluster each data vector belongs to for any specified number of clusters.
- Agglomerative hierarchical cluster analysis starts with as many clusters as data vectors. Data vectors are successively agglomerated into clusters until only one data vector remains. The result of this can be shown in a dendrogram. The dendrogram is the tree-like diagram that can show the data vectors, which have been clustered at each agglomeration sequence.
- Often, but not always, different agglomerative clustering methods for analysing data can yield different results. In particular for small data sets, different methods might produce similar results.

IV. Acknowledgments

The author wishes to thank all those who dedicated their time answering my queries and providing me with valuable comments during the preparation of this study.

Conflicts of Interest

The author declares no conflict of interest.

References Références Referencias

1. Hermann Moisl. *Cluster analysis for corpus linguistics*. De Gruyter Mouton: Berlin, 2015.
2. Everitt, B. S.; Landau, S.; and Leese, M. *Cluster Analysis*. 4th ed.; Arnold: London, 2001.
3. Everitt, B. S. Unresolved Problems in Cluster Analysis. *Biometrics*1979, 35, 1, 169-181.
4. Charles Romesburg. *Cluster Analysis for Researchers*. Wadsworth Inc: USA, 1984.
5. Anil Jain and Richard Dubes, R. *Algorithms for clustering data*. Prentice-Hall Englewood Cliffs: NJ, USA. 1988.
6. Michael Steinbach. Available online: http://www-users.cs.umn.edu/~han/dmclass/cluster_survey_10_02_00.pdf (accessed on 20 September 2015).
7. Patricia Pulliam Philips and Cathy A. Stawarski. *Data Collection Planning and Collecting.* All Types of Data, Pfeiffer: USA, 2008.
8. Craig J. Cleaveland. *An Introduction to Data Types*. Addison-Wesley Longman, Incorporated: Michigan, 1986.

© 2016 Global Journals Inc. (US)

9. Pyle D. *Data preparation for data mining*. CA: Morgan Kaufmann Publishers: San Francisco, 1999.
10. Michael Anderberg. *Cluster analysis for applications*. Academic Press, Inc: London, 1973.
11. Coxeter H. S. M. *Non-Euclidean Geometry*. The Mathematical Association of America: USA, 1998.
12. Berry Bonola. *Non-Euclidean Geometry*. Lighting Source: USA, 2007.
13. Deza, M. and Deza, E. *Encyclopedia of Distances*. Springer: Berlin, 2009.
14. Rui Xu and Donald C. Wunsch. *Clustering*. Wiley-IEEE Press: Hoboken, New Jersey, 2009.
15. Baker Kirk. Available online: https://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf (accessed on 10 September 2015).
16. Rencher, A. C. and Christensen, W. F. *Methods of multivariate analysis*. 3rd ed., John Wiley and Sons In: Hoboken, New Jersey, 2012.
17. Amit Singhal. Available online: http://singhal.info/-ieee2001.pdf (accessed on 2 July 2015).
18. Lance, G. N. and Williams, W. T. A general theory of classificatory sorting strategies I. hierarchical systems. *Computer journal*1967, 9, 373-80.
19. Milligan, G. W. A study of the beta-flexible clustering method. *Multivariate Behavioral Research* 1989, 24, 163-176.
20. Berry, M. W. and Browne, M. *Lecture notes in data mining*. Fu Island offset printing: Singapore, 2006.
21. Rosie Cornish. Mathematics Learning Centre. Available online: http://www.statstutor.ac.uk/resources/uploaded/clusteranalysis.pdf (accessed on 22 July 2015).

© 2016 Global Journals Inc. (US)

Global Journal of human-social science: G
Linguistics & Education
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-460X & Print ISSN: 0975-587X

Role of a Teacher in Teaching Speaking by Following a Communicative Approach: To What Extent is this Possible in an ESL Context like Bangladesh?

By Shafinaz Sikder
BRAC University, Bangladesh

Abstract- This paper is predominantlya short research which investigates the issue of how a teacher can promote the practice of speaking English by following a communicative approach within the traditional context of Bangladeshi language classrooms. The paper thus consists of four segments of literature reviews based on some secondary data collected from contemporary online journals and books. In a brief overview, it can be said that the paper starts with an assessment of the role of teachers in case of prioritizing fluency over accuracy in our language classrooms. This is followed by a discussion on the common sorts of speaking activities which an ESL teacher is ought to follow to promote fluent speaking within the classroom contexts. After this, the second last section focuses on some basic problems which act as hindrances in case of teaching speaking communicatively; for example the negative impact of Bangladesh's language based identity and its influence on teaching English Speaking. Lastly the paper sheds light on the controversial issue of whether it is even possible to teach speaking communicatively to our young school learners or not; and if so than to what extent is it actually possible.

Keywords: communicative approach, speaking activities, ESL speaking context, role of a teacher.

GJHSS-G Classification : FOR Code: 930299

Strictly as per the compliance and regulations of:

© 2016. Shafinaz Sikder. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http:// creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Role of a Teacher in Teaching Speaking by Following a Communicative Approach: To What Extent is this Possible in an ESL Context like Bangladesh?

Shafinaz Sikder

Abstract- This paper is predominantlya short research which investigates the issue of how a teacher can promote the practice of speaking English by following a communicative approach within the traditional context of Bangladeshi language classrooms. The paper thus consists of four segments of literature reviews based on some secondary data collected from contemporary online journals and books. In a brief overview, it can be said that the paper starts with an assessment of the role of teachers in case of prioritizing fluency over accuracy in our language classrooms. This is followed by a discussion on the common sorts of speaking activities which an ESL teacher is ought to follow to promote fluent speaking within the classroom contexts. After this, the second last section focuses on some basic problems which act as hindrances in case of teaching speaking communicatively; for example the negative impact of Bangladesh's language based identity and its influence on teaching English Speaking. Lastly the paper sheds light on the controversial issue of whether it is even possible to teach speaking communicatively to our young school learners or not; and if so than to what extent is it actually possible. To be more specific about the goal of this paper, it can be said that it is trying to grab our language teachers' attention towards the fact that they can play a very role promote in case of teaching speaking communicatively within the existing system. In other words, the paper is actually restating the fact again and again that it's a "teacher" who can bring about a lot of communicative changes in the techniques and methods of teaching speaking even within these traditional classrooms if and only he/she wants.

Keywords: communicative approach, speaking activities, ESL speaking context, role of a teacher.

I. Introduction

a) Introducing Speaking and its Importance

Why different schools of South Asia are promoting English teaching in the recent years? What could be their main purpose behind teaching English to their young learners? Is it just for the young kids to know English by heart? Or is it to gain practical interactional skills in the language? If the latter one is the main motive, then why the young learners of the South Asian countries including our country are studying English by means of traditional methods? (Haider & Chowdhury, 2012, p.13). Usually any language learner will first have an interest to speak and therefore function successfully in the language. However, in this part of the world, especially in Bangladesh, learning English to speak and function could sound quite challenging over the last few decades. This is because in our country, teaching of speaking is not done in the way Nunan, Ara and other ELT specialists has suggested. As explained by Nunan (2003), teaching speaking could include not only learning speech sounds/ patterns, use of words and sentence stress, intonation patterns, use of appropriate words and sentences according to the proper social setting, audience, situation and subject matter but also the fluent use of language as a means of expressing values and judgments(p. 330).

b) Research Purpose

Therefore the main purpose of this paper is to focus on how teachers can teach speaking fluently by means of communicative approach within the Bangladeshi language classrooms. Thus a Bangladeshi perspective has been under taken in this research work to analyze the following issues: What can be some main barriers that are actually keeping the students away from speaking English fluently? What happens in the classrooms where speaking is taught to the young learners? And why is it being considered less important than the other productive skill (writing)?

As stated in a current research done in Bangladesh, it can be said that "this is not the right way of teaching speaking" (Ara, 2009, p. 167). According to Ara (2009), students should learn speaking English by means of fun activities since some traditional educational approach has prevented learning rather than facilitating it (p.167). Moreover, it is a country where both formal and informal speaking skills are being taught by following the most common traditional methods since its independence (Imam, 2005, p.474). The truth is, most of the primary sections of different schools (both Bangla and English medium) are practicing traditional pedagogy where students are

Author: Department of English and Humanities, BRAC University, Dhaka, Qamrun Noor Apartment, Sir Sayeed Road, Mohammadpur, Dhaka, Bangladesh. e-mail: shafinaz_sikder@hotmail.com

© 2016 Global Journals Inc. (US)

nothing more than passive recipients (Haider & Chowdhury, 2012, p.13). Actually something that has been always missing in this context, is the practice of teaching speaking fluently with the help of different 'free activities'[1] like information gaps, role plays, simulations, games, brainstorming, storytelling, picture describing etc.(Gower, Phillips & Walters, 1995, p. 99-110). So basically, this study consists of works of some language researchers who tried to emphasize on how teaching speaking can be done in a much more communicative[2] way even within an existing traditional setting.

c) Research Statement

Teachers can bring about a lot of communicative changes in case of teaching speaking within the existing traditional classrooms.

d) Research focal questions

Some of the questions which have been addressed in the paper are:

- What should a teacher focus on in a primary level: accuracy or fluency?
- What can be some suggested activities related to teaching speaking communicatively? Therefore can our teachers follow them to promote fluent speaking?
- What role is being played by Bangladesh's "language based identity" in case of teaching English with much importance?

e) Research Significance

The significance of this study lies in creating awareness among all the ESL teachers about the fact that they actually have a very crucial role to play whether they acknowledge it or not. That is to say, our teachers can be proved to be some important personalities in case of undertaking a communicative way of teaching speaking if they want to. Instead of walking into a class and giving out a lecture, they can actually involve the students into some meaningful speaking practice. They can be real pioneers in case of altering the traditional way of practicing speaking. The fact is, they can with time, bring about a tremendous change in our entire teaching approach if they start taking attempts from now on.

f) Research Limitations

The most noticeable limitation of this paper is that the research has only been conducted over a few secondary sources; overlooking a practical investigation of conducting a survey on the discussed issue. That is to say no practical implementation of the theories has been carried out to gather quantitative/qualitative data from primary sources. Besides, the study has considered only the major number of schools whereas there are a few schools which are teaching speaking with primary importance now-a-days.

II. Literature Review

a) What should be a Language Teacher's Concentration: Accuracy or Fluency?

In 1995, Gower, Phillips and Walters explained that speaking can be learned by following two basic ways: accuracy or fluency. Accuracy would mean learning of the correct use of vocabulary and grammar items by means of controlled and guided activities while fluency would mean the ability to get the message across regardless any grammatical mistakes (p. 100). As Gower et al. (1995) advised, the students should not be corrected during their speaking since the natural use of incomplete sentences, common expressions, use of filters, hesitations etc. might be hampered (p. 100). Gower et al. (1995) also advised that a teacher should not interrupt and give correction to students during their speaking tasks (p. 100). They added that in any particular activity, the teacher can make it clear that in which areas, accuracy is expected and to what extent, so that the students' anxiety to 'get it right' does not interfere too much with their fluency and ability to communicate (p. 100).

Therefore the question arises, how our teachers can promote speaking with fluency? Is promotion of fluent speaking possible with the current traditional approach? Or it demands a shift in teaching approach? The following section will shed light on this matter.

b) Activities Promoting Fluent Speaking: Undertaking a Communicative Approach

Gower, Phillips & Walters (1995) said that teachers can promote fluent speaking by means of different free/creative activities; it mean spractice of fluency by using real life language in different real situations etc. (p.100-111). Another communicative activity is the 'Information gap activity' where students complete gaps in pairs by means of communications (p. 103). Besides, 'role plays' where students assumes a particular person's character to act out a conversation can also be very effective in most of the contexts (p. 105). Apart from these, 'simulation' where students being a particular person solves a task in a given situation is also helpful although it can be lengthy and time consuming (p. 107). The common most activity is 'group discussions' where classroom talking takes place among students and they share their ideas by using particular language structures and vocabulary without the teacher's interruption (p. 107). Other communicative activities which teachers can practice in the classroom are different "interactive games" or "picture descriptions tasks" along with "storytelling activities" where students

[1] Free activities refer to activities which promotes creative practice opportunities for general fluency practices (Gower, Phillips & Walters, 1995, p. 100-110)

[2] Communicative ways means teaching speaking in an interesting way by fun activities which promotes lots of interaction among the students. (Gower, Phillips & Walters, 1995, p.101)

© 2016 Global Journals Inc. (US)

are required to tell imaginative stories in given contexts (Murphy, 1991, p. 55). Apart from Murphy, Scrivener also suggested some communicative activities like "picture difference tasks", "group planning tasks", "ranking tasks", "pyramid discussion", "board games", "puzzles and problems", "real play" etc. which can actually help student speak up with whatever resource they have (p. 218-220).

Now the problem lies somewhere else. That is to say, all the suggested activities should have been easily used as interesting prompts by our teachers and therefore also accepted with consent among our students if and only their mindset towards learning English could have been changed. To be more specific, learning English as a "functioning" language is not something very positively taken by many of our students and teachers since we have a language based identity as a nation; this issue has been briefly discussed in the subsequent section.

c) *Impact of Bangladesh's Language Based Identity on Its Teachers and Teaching Methods*

In 2005, Imam explained the issue of Bangladeshi's unique history of language movement and how it has brought about independence for the nation. He linked this crucial issue to Bangladeshi peoples' sentiments and emotions (p.471). In his words, "One problem is that, in the minds of most of the people, national identity and speaking English are positioned as antagonistic, not complementary" (p. 471). He added that since the declaration of independence of 1971, Bangla was confirmed as the medium of instruction by our teachers in the governmental schools (p. 474). He also added that "Bangladesh, being a new country and having its origin in the glorious 1952 language movement, is comparatively new in this English language speaking promotion race; moreover as a nation it has also been shaped by language and the politics of language, to an unusual degree" (p.472). He concluded his paper by saying that "The global English might function as a Trojan horse: as a displacer of national tradition and an instrument of continuing imperialist intervention which has actually kept away the teachers from promoting English speaking in a true sense" (p.472).

Still with all these problems related to our national history and language based identity, it is a teacher, who can bring upon necessary changes. This is actually the most important sort of remedy in case of solving our problem.

d) *Teachers: Can they make it?*

Imam (2005) stated that a teacher can bring about a lot of significant changes within the classroom if he/she wishes. In fact, the way a student will learn to speak depends on how he/she is being taught (p. 475). This also depends on a teacher's perspective and personal variables like motivation, interest, experience, attitude towards job etc. Regarding the role of the teacher, Scrivener also (1994) mentioned about some of the communicative techniques that a teacher can use in case of teaching speaking fluently; scaffolding is one of them by which teachers can provide spontaneous correction without interfering much and by helping the speaker to construct his/her conversation (p. 227). Scrivener also suggested that scaffolding can be of great help if it is done by showing interests by nodding, making eye contacts, asking for clarifications of unclear information, by encouraging echo, asking conversation-oiling questions etc. (p. 227). Besides there are some other aspects which can be accelerated by teachers like lessening teacher talk time, eliciting ideas rather than lecturing, involving the students into argumentative topics of discussions, providing enough scope for student-student interactions and creating non-threatening situations where students can communicate comfortably etc. (Ferris & Tagg, 1996, p. 301).

III. Overall Discussions

From the above study, it is evident that teaching speaking with a communicative approach seems quite difficult in real classrooms since speaking as a skill has always received less importance in our context where students never have to take oral tests. The fact is, since the students do not have to deal with speaking exams, the teachers and therefore the whole educational system did never emphasize teaching speaking with importance.

Besides other researchers, Haider and Chowdhury (2012), showed how teaching speaking needs more emphasis in our country. They said, "Despite the introduction of a communicative syllabus, NCTB had the realization that a syllabus alone could not ensure the implementation of communicative language teaching and learning in the classroom; it could only provide a set of criteria for a change. Therefore, in order to get the best out of the new syllabus a number of reform initiatives such as preparing suitable teaching materials, arranging teacher training and bringing about changes in the examination system were recommended." (p. 12).

Moreover, Ara (2009) said, "little improvement has taken place in the overall English proficiency among the students" (p. 167). This is only because of inaccurate teaching techniques of English towards the young learners. Therefore in case of bringing changes, the first sort of alterations should be brought upon by teachers who teach the young ones (p.167). In other words, the teacher must try to initiate necessary changes among the existing teaching method, approach, procedure, activities, classroom arrangement and management system. This however, will not be accepted easily by the students or the school authorities. Even if the school authority takes this into consideration yet there is a long way to go. This is

© 2016 Global Journals Inc. (US)

because sudden changes can be easily made in theories but practical implications take time. Nevertheless, changes in real sense can be made possible over a long time. This is because bringing some changes (in teaching process) little by little is neither impossible nor unacceptable.

IV. Conclusion

a) A Few Recommendations

A teacher should be more than a teacher; he/she should be a facilitator. The truth is, language learning is a process that depends a lot on how a teacher teaches. As Harmer (2007) said, a teacher can act in many ways: as a controller, prompter, participator, resource or as a tutor (p. 108). Unfortunately, in our context, the teacher is the controller of a classroom. Incase of his suggestions of facilitate learning, he said that the teacher should be democratic and should let his/her students participate in decision makings rather than controlling the class (p.107). That is to say, teachers must allow learners' autonomy in classrooms. Besides, Ur (1996) also emphasized a teacher's role in case of teaching speaking. He said that teachers can do a lot within the class to solve speaking problems (p. 121). They can promote activities like (1) use of group/pair works, (2) make a careful choice of topic and task to stimulate interests and (3) keep students speaking the target language no matter what (p.121).

Scrivener (1994) also said the same thing. He also stated that a teacher has a lot to do to promote learning in a classroom (p. 228) He also mentioned a few suggestions which can help a teacher prepare speaking lessons better like the use of "structuring talk" (making sure that learners get maximum chance to talk in the class by reducing teacher's participation level), avoiding "talk-talk loop" (to avoid asking several questions to which learners cannot response instead of asking one single question and shutting up), using "open questions" (asking questions whose answers requires more than just 'yes' or 'no' from the students' part) etc. (p. 228).

To sum up the above discussion, it can be said that the role of a teacher is very much important in case of teaching speaking communicatively. This is because it is the teacher who can actually make his/her students speak fluently by incorporating a few changes with his/her innovative use of methods, techniques and activities within the classroom.

V. Acknowlegment

To start with, I would like to show my gratitude towards my parents. After them, my special thanks go to my course instructor M.D. Mahmudul Haque. His advice, guidance, suggestions, feedbacks and patience were the driving forces for me.

References Références Referencias

1. Ara, Shaheen. (2009). Use of Songs, Rhymes and Games in Teaching English to Young Learners in Bangladesh. *The Dhaka University Journal of Linguistics*, *2*(3), 161-172. Retrieved from http://www.banglajol.info/index.php/DUJL/article/-viewFile/4150/3439
2. Ferris, D. & Tagg, T. (1996). Academic Listening/ Speaking Tasks for ESL Students: Problems, Suggestions, and Implications. *TESOL Quarterly*, *30*(2), 297-320. Retrieved from http://www.jstor.org/-stable/3588145
3. Gower, R., Phillips, D., Walters, S. (1995). *Teaching Practice Handbook (*2nd ed.). United Kingdom, UK: Macmillan Heinemann ELT.
4. Haider, M. Z. & Chowdhury, T. A. (2012). Repositioning of CLT from Curriculum to Classroom: A Review of the English Language Instructions at Bangladeshi Secondary Schools. *International Journal of English Linguistics*, *2*(4), 12-22. Retrieved from http://dx.doi.org/10.5539/ijel.v2n4p12
5. Harmer, J. (2007). *The Practice of English Language Teaching.* (4th ed.). London: Pearson Longman.
6. Imam, S. R. (2005). English as a Global Language and the Question of Nation-Building Education in Bangladesh. *Comparative Education, 41*(4), 471-486. Retrieved From http://www.jstor.org/-stable/30044556
7. Murphy, J. M. (1991). Oral Communication in TESOL: Integrating Speaking, Listening, and Pronunciation. *TESOL Quarterly, 25*(1), 51-75. Retrieved from http://www.jstor.org/stable/3587028
8. Nunan, D. (2003). *Practical English Language Teaching*. New York, NY: McGraw-Hill Book.
9. Scrivener, J. (1994). *Learning Teaching: The Essential Guide to English Language Teaching* (3rd ed.). Macmillan Publishers Limited.
10. Ur, P. (1996). *A Course in Language Teaching: Practice and Theory.* United Kingdom, UK: Cambridge University Press.

© 2016 Global Journals Inc. (US)

Global Journal of human-social science: G
Linguistics & Education
Volume 16 Issue 3 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-460X & Print ISSN: 0975-587X

Agglomerative Hierarchical Clustering: An Introduction to Essentials. (3) Standardization, Normalization, and Dimensionality Reduction of a Data Matrix

By Refat Aljumily
University of Newcastle, United Kingdom

Abstract- In a previous tutorial article I looked at a proximity coefficient and, in the light of that proximity created a vector-distance matrix and used it to construct a hierarchical tree using different hierarchical clustering methods which will be the basis for exploratory multivariate analysis. The present article deals with three topics: (i) standardization for variable scales variation, (ii) normalization for sample length variation, and (iii) dimensionality reduction or minimization of data space. These techniques reflect the author's academic background and particular area of interest and are, by necessity, not a particular purpose and are straightforwardly applicable to other kinds of data, and thus to a wide range of analysis in Linguistics. My treatment of these techniques is, necessarily, introductory and brief. I hope that this article will provide practitioners with an introductory overview of these techniques used for cluster analysis of electronic corpora of linguistic data.

Keywords: corpus, vector, matrix, standardization, coefficient of variation, normalization, dimensionality reduction.

GJHSS-G Classification : FOR Code: 139999

AGGLOMERATIVEHIERARCHICALCLUSTERINGANINTRODUCTIONTOESSENTIALS3STANDARDIZATIONNORMALIZATIONANDDIMENSIONALITYREDUCTIONOFADATAMATRIX

Strictly as per the compliance and regulations of:

Research | Diversity | Ethics

© 2016. Refat Aljumily. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http:// creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Agglomerative Hierarchical Clustering: An Introduction to Essentials. (3) Standardization, Normalization, and Dimensionality Reduction of a Data Matrix

Refat Aljumily

Abstract- In a previous tutorial article I looked at a proximity coefficient and, in the light of that proximity created a vector-distance matrix and used it to construct a hierarchical tree using different hierarchical clustering methods which will be the basis for exploratory multivariate analysis. The present article deals with three topics: (i) standardization for variable scales variation, (ii) normalization for sample length variation, and (iii) dimensionality reduction or minimization of data space. These techniques reflect the author's academic background and particular area of interest and are, by necessity, not a particular purpose and are straightforwardly applicable to other kinds of data, and thus to a wide range of analysis in Linguistics. My treatment of these techniques is, necessarily, introductory and brief. I hope that this article will provide practitioners with an introductory overview of these techniques used for cluster analysis of electronic corpora of linguistic data. The assumption is that the data is in the form of an *m* x *n* matrix D in which, may require to transform it in various ways prior to cluster analyzing it. Standardized data matrix enables practitioners to measure the variation between n-variables and to cluster the cases they describe in common scales and values, regardless of their original scales and values. Normalized data matrix enables practitioners to eliminate the effect of variation in length among n-samples and to cluster them as if they were all (about) the same length, regardless of their original length. Dimensionality-reduced space data matrix enables practitioners to select and/or extract *n*-most interesting variables relevant to the research question and to visualize an existing pattern, regardless of the original space. A worked example is given to illustrate the effect each transformation technique has on a given data matrix. These transformation techniques have their own strengths and weakness but are beyond the scope of my objectives in this article.

Keywords: corpus, vector, matrix, standardization, coefficient of variation, normalization, dimensionality reduction.

I. Introduction

Language corpus typically consists of more or less numerous texts each of which is described in terms of the selected linguistic features, technically known as variables. If it is to be analyzed using clustering methods, the selected variables need to be mathematically represented. A widely used way of doing this is vector space representation. Where vector space representation is used, each text is described by a vector, and the language corpus is consequently a set of vectors. Such a set of vectors is conveniently represented as a matrix in which the rows are the texts and the columns the linguistic features (variables). Thus, language corpus consisting of *m* texts each of which is described by *n* variables is represented by an *m* x *n* matrix D in which D_i (for $i = 1...m$) is the *i*'th text, D_j (for $j = 1..n$) is the *j*'th variable, and D_{ij} the value of variable *j* for text *i*. Once the language corpus has been constructed in a matrix, it is important to consider the issues relevant to cluster analysis of texts. Three types of issues are considered: (i) variable scales variation, (ii) text length variation, and (iii) variables selection/ extraction. This article proposes ways to remove the effect of each of these issues: (i) normalization for variation in text length, (ii) standardization for variation in variable scales, and (iii) dimensionality reduction. These techniques can be used, if it is necessary, to transform a given data matrix prior to analyzing it.

II. Transformation Techniques

a) Variation of variable scales

Almost any linguistic feature in a corpus such as word-forms, sentences, grammatical sequences, parts of speech, or any other easy to count features, can be measured. We use measurements to examine these linguistic features mathematically. In general, when we measure a linguistic feature, we define or interpret its properties in relation to special scales or units of measurement, then recording its happenings. That measurement constitutes the values of the linguistic features, for example: function words usage= 3000, average word-length=3, number of punctuation marks=500, diversity of words in a text 10%, and so on. Measurement is fundamental in the creation of language data because it makes a link between a particular linguistic feature in mind and an activity that originates from an individual, and thus allows the results of cluster analysis to generate a hypothesis about a language or language user. Measurement is only possible in terms of

Author: University of Newcastle, Ashfield close, Newcastle Upon Tyne. e-mail: refat.aljumily@newcastle.ac.uk

© 2016 Global Journals Inc. (US)

some scale. Scales are systems designed to tell us how much of a measurable characteristic a given variable has. Scales have different types of numerical units and ranges (scales of measurements) appropriate to them which carry different amounts of information in any given application. The variables selected for describing linguistic features involving cluster analysis may require measurement on different scalars. If variables are measured on different scales, variables with large values contribute more to the distance measure than variables with small values.

Given an *m* x *n* data matrix M in which the *m* rows represent the *m* objects to be clustered, the *n* columns represent the *n* variables, and the entry at M*ij* (for $i = 1..m$, $j = 1..n$) represents a numerical measure of object *i* in terms of variable *j,* a clustering method has no idea what the values in the data matrix mean and calculates the degrees of similarity: variables that are measured in large values will have a greater influence on the degrees of similarity between the objects than those variables measured in smaller values, and, therefore, will affect the reliability of the cluster analysis. To see this, take a look at the following data matrix which describes nine students (A, B, C, D, E, F, G, H, I) in terms of their use of three linguistic features in the academic papers, one of which represents the total number of contractions, another one function word/content word ratio, and a third function words frequency.

Table 1 : A data matrix with different variable scales

Students	Number of contractions	FW/CW (percentage)	FW (frequency)
A	187	40	27000
B	185	35	25000
C	184	33	26000
D	170	29	23500
E	166	25	22000
F	164	26	21000
G	160	60	15000
H	150	53	10000
I	159	61	14500

In Table/1the first column variable represents the total number of contractions, the second FW/CW ratio in percentage, and the third FW in frequency. A hierarchical cluster analysis of the matrix rows using Squared Euclidean distance gives the following dendrogram:

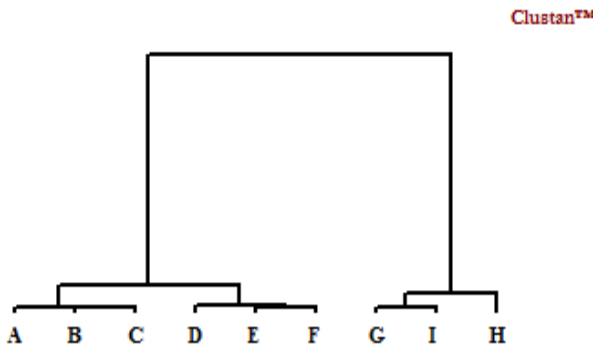


Figure 1 : Hierarchical clustering of 9 students based on different linguistic features measured on different scales

In Table/1 the largest values are those in the function words column, and the corresponding agglomerative clustering dendrogram in Figure/1 classifies the students into three main clusters (27000-26000), (23000-21000), and (10000-14500) by function words. In other words, the clustering analysis didn't find any significant clusters; there is a clear and very strong tendency to cluster by scale of measurement. The essence of the problem now is that we need a clustering structure that reveals the proximities among the vectors independent of the variation in scaling.However, there are many standardization methods as a technique for removing the effect variation in scaling among data and making each variable receives equal contribution in the cluster analysis. Some of these methods are:

- Standard or Z-score standardization method.
- Standardization method based on variable mean.
- Standardization method based on variable sum.
- Cosine standardization method.
- Max standardization method.
- Range standardization method.

One of the reasons for this diversity is that different standardization methods are required for different purposes; for clustering or for other purposes. No one single standardization method will be suitable for all applications. Some methods can be extremely useful even if they are mathematically limited. Other methods bring different benefits, although some bring disadvantages as well. To be suitable for cluster analysis, however, a method must preserve differences in variability among variables, thereby giving a true account of the intrinsic cluster structure of the unstandardized data matrix. The emphasis is the degree to which a method preserves the pre-standardization intrinsic variabilities of variables in post standardization

© 2016 Global Journals Inc. (US)

absolute magnitudes of variability. By the intrinsic variability, we mean the amount of variability in the values of a variable expressed independently of the scale of those values and measured in statistics by the coefficient of variation, which is defined with respect to a variable *v* as the ratio of *v*'s standard deviation to its mean, and by the absolute magnitude of variability we mean the amount of variation in the values of a variable expressed in terms of the scale of those values, and is measured by the standard deviation.

A standardization method based on variable means does this in the sense that it has the effect of preserving intrinsic variability in the values of a variable, and it does that in the following way: individual numerical column vectors of unstandardized data matrix can be standardized in relation to their mean, where the value of a given numerical column vector - V in the unstandardized matrix must be divided by the mean μV of column vectors:

V_i std $= V_i / \mu V$

Where:

- V_i std is a standardized column vector in a data matrix, for $i=$ 1…number of rows in matrix or, equivalently, the number of text files in a corpus.
- V*i* is an unnormalized document vector, for *i* as above.
- μV is the column vector mean, or scalar, measured by the total number of values in each column vector.

To illustrate this, the first three students described by the total number of contractions, FW/CW ratio (in percentage), and FW (in frequency), in the data matrix of Table/1 are recalculated.

Table 2 : MEAN standardization of the matrix in Table/1

students	Contraction	FW/CW	FW	Contraction	FW/CW	FW
A	187	40	27000	1.01	1.11	1.03
B	185	35	25000	1	0.97	0.96
C	184	33	26000	0.99	0.91	1
Std	1.247	2.943	816.496	0.084	0.022	0.028
CV	0.006	0.081	0.0314	0.084	0.022	0.028
	a. unSTD matrix of Table (1)			b. Mean STD matrix of Table (1)		

In Table/2, it is clear that MEAN-standardization has made the variation magnitudes comparable and also has preserved the coefficients of variation of the unstandardized variables. This is because division by a scalar, here the column vector mean, is a linear operation that alters the scale while preserving the shape of the original value distribution. It is also clear that the standard deviations of contractions, FW/CW ratio, and FW in Table 1b are identical to the corresponding coefficients of variation. This is because, for any data vector (here representing persons), it is always the case that its coefficient of variation is identical to the standard deviation of the MEAN-standardized version of vector. After standardizing the variables for the remaining persons as above, the application of a hierarchical method on the standardized data matrix in Table 1b shows sufficiently accurate clustering; the hierarchical tree in Figure/2 differs substantially, and it clusters the nine students according to the relative magnitude of values in the matrix columns, i.e. regardless of the variation in the variable scales.

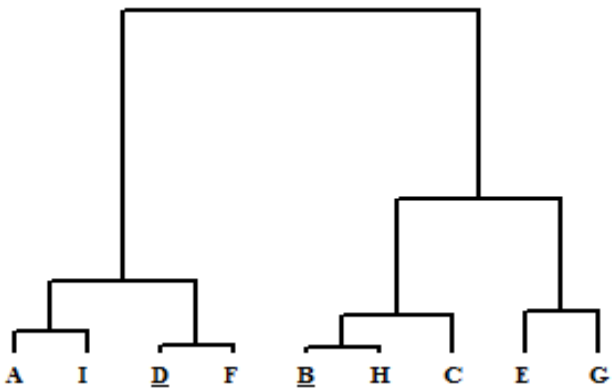


Figure 2 : Hierarchical clustering of the standardized data matrix in Table/1

For more on this technique see, for example, [Moisl, 2015; Chu, Holliday, and Willett 2009; Gnanandesikan, Tsao, and Kettenring 1995; Milligan and Cooper 1988].

b) Normalization for variation in sample length

A corpus is a collection of texts collected with a particular linguistic research project. Very often, it happens that a corpus contains texts of varying sizes; many of them can be disparate in length and not at all identical with each other. If the disparity varies greatly from text to text, a critical issue arises that must be taken into account: the data abstracted from the corpus for cluster analysis will give distorted results and

© 2016 Global Journals Inc. (US)

consequently it becomes difficult to accurately indicate much in terms of similarities, or differences, between the texts. To see the effect of length variation on clustering performance, an agglomerative hierarchical analysis of a corpus consisting of some varying-length texts is carried out and the result is shown in Figure/3:

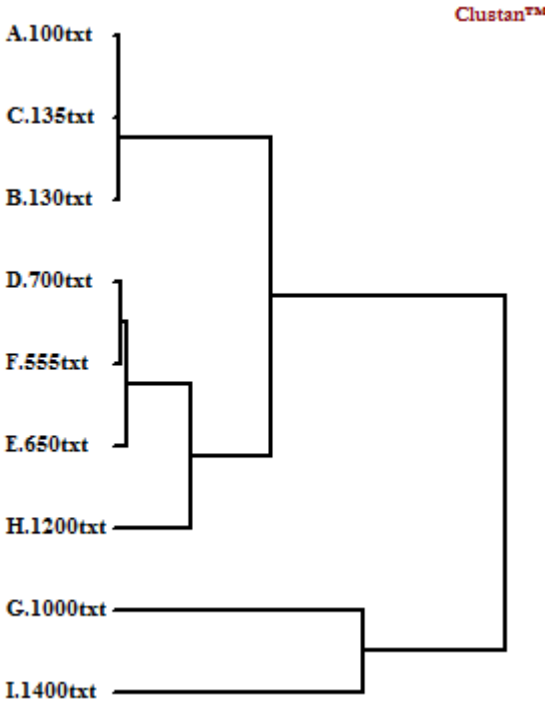


Figure 3 : Clustering based on the text lengths (prior to length normalization)

In this figure, there is a progression from the shortest texts at the top of the tree to the longest at the bottom and this means that there is a clear and very strong tendency to cluster by length. This can easily be seen from the number to the right of each of the text names which represents the number of words in the text. The reason for this is that, in the present example, the data abstracted from a corpus is based on frequency; each vector contains frequencies of lexical types for one of the texts, and a set of vectors are stored as the rows of the data matrix. In this sense, variations in the row vector lengths are simply a result of variations in magnitudes of lexical frequencies stored on the data matrix row vectors. To understand this, assume counting the number of occurrences of some lexical type *j* in a corpus containing two texts, A and B. Assume that *j* occurs 10 times equally across those two texts. After entering the lexical frequencies into data matrix row vectors, the interpretation would obviously suggest that on the basis of their usage of *j*, the two texts A and B are identical and that *j* apparently fails to discriminate between text A from text B. If, however, one knows that text A is 5000 words long and text B 500 words long, this is no longer the case. It is clear that, although both texts have the same frequency of occurrences of *j*, its significance level in them is significantly different from each other. The lexical type *j* is relatively infrequent in text A and relatively frequent in text B and therefore this difference can be used to differentiate between those texts. If we assume again that the text B is 50000 words long instead of 500, based on its observed frequency in 500 words, then there would have been 1000 occurrences of *j*. In short, the longer a text, the more likely in general a given word with a specific probability of occurrence is to occur in it, and, if it occurs, the higher the frequency of occurrence is in general likely to be. These different text lengths, called variations in lengths, are inherent in all texts in collections and result in variations in the frequencies stored in the data matrix. The variation may be large or very small, but it is always present. For the cluster analysis to be accurate and reliable, weighting to compensate for variation in text length is therefore necessary to remove this effect. The common way to do so is to adjust the data matrix so that not just frequency but its significance relative to text length can be represented and thus incorporated into subsequent analysis. There are a number of normalization methods that are theoretically motivated, for example:

- cosine normalization
- probability normalization
- normalization by mean term frequency within document
- normalization by maximum term frequency within document
- normalization by mean document length across collection
- normalization by maximum document length across collection.

but, the one most easy to understand is normalization by the mean document length across collection, and the reminder of discussion will concentrate on that. In this method, to adjust the lengths of each row vector of an m × n data matrix of lexical types frequencies, the frequency count for a given lexical type in a given text must be multiplied by the mean length of all texts then divided by the total number of frequency counts occurring in that text. The effect of this process: decreasing the values in the vectors that represent long texts, increasing them in vectors that represent short ones, and, for texts that are near or at the mean, to change the corresponding vectors little or not at all. This can be expressed as:

$$X'i = x_i \frac{\mu}{lengt\ h_i}$$

where X here in relation to mean length of texts in a corpus:

© 2016 Global Journals Inc. (US)

- X'i is the normalized frequency of *i*' th lexical type in a row vector, for *i*=1…….n.
- Xi is unnormalized frequency of *i*' th lexical type in a row vector.
- μ is the mean length of vectors across all texts (T). This obtained by dividing the sum of frequencies of matrix row vectors (T) by that of the number of texts *n*, for *i*=1…..*n*:

$$\mu(T) = \frac{\sum i = 1 \ldots n \; length_{T_i}}{n_{Ti}}$$

- Length (i) is the sum of frequencies of any row vector (i).

For example, let M below be a matrix having 3 texts (a, b, c) with unnormalized values of four lexical types as shown below:

	V1 the	V2 a	V3 you	V4 I
txt.a (length= 500)	12	15	3	53
txt.b (length=1500)	4	36	1	36
txt.c (length=2430)	7	80	0	29

using the formulas above:

- we need to find the mean length across all texts. Thus we have 500 +1500+2430 / 3 = 1476
- in each row vector, the count for a given lexical type is multiplied by the mean text length, then divided by the total number of frequency counts occurring in that row vector. Thus, we obtain:

For txt.a we have:	For txt.b we have:	For txt.c we have:
12×(1476/500)=35.42	4×(1476/1500)= 3.93	7×(1476/2430)= 4.25
15× (1476/500=44.28	36×(1476/1500)= 35.42	80×(1476/2430)=48.59
3×(1476/500)=8.85	1×(1476/1500)=0.98	0×(1476/2430)=0
53×(1476/500)=156.45	36×(1476/1500)=35.42	29×(1476/2430)=17.61

This way the resulting normalized matrix looks like:

	V1 the	V2 a	V3 you	V4 I
txt.a (length= 500)	35.42	44.28	8.85	156.45
txt.b (length=1500)	3.93	35.42	0.98	35.42
txt.c (length= 2430)	4.25	48.59	0	17.61

The effect of the normalization method on the data matrix shown in this example above is clear: all the values in txt.a have been substantially increased because it is significantly shorter than the mean text length: length-500 <1476 (the mean). For txt.b, the values have been slightly decreased because it is slightly longer than the average document length: length-1500 >1476. Finally, the values for txt.c have been substantially decreased because it is significantly longer than the average document length: 2430> 1476.

Applying this to the example in Figure/3above, an agglomerative hierarchical tree of the normalized data matrix row vectors is shown below, where clustering by relative magnitude of values in the matrix rows is now in evidence.

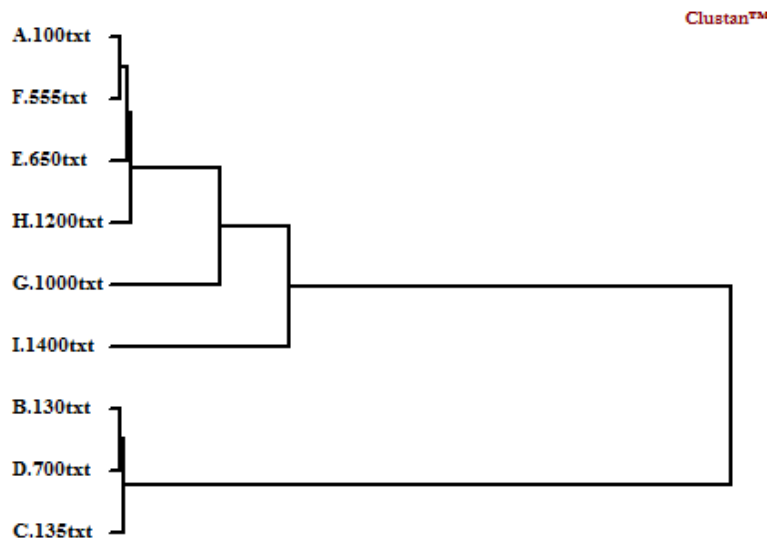


Figure 4 : Clustering based on the normalized matrix row values

© 2016 Global Journals Inc. (US)

In summary, normalization enables us to cluster and compare texts with each other irrespective of their lengths and failure to normalize for variation in text length can produce fundamentally erroneous cluster analytical results. Nevertheless, the process of normalizing data matrix column or row vectors itself has some unresolved problems and these problems are not discussed here. More on document length normalization can be found in, e.g., [Moisl, 2015; Priddy and Keller, 2005; Belew, 2000; Singhal et al., 1995 and 1996].

c) Dimensionality reduction

Dimensionality is a major issue for data analysis in any given application. Where the aim is to generate a matrix M in which the rows are the data points, the column variables are lexical types, and the value at any given matrix location M_{ij} is the frequency of lexical type *j* in *i*, dimensionality has a particular relevance to the application of cluster analysis. In dealing with high-dimensional data, however, having too much is rarely a problem. Quite the opposite --the usual situation with high-dimensional data is that there is far too little. High-dimensional spaces are inherently sparse, and, to achieve adequate definition of the data manifold, the amount of data required very rapidly becomes intractably large; this phenomenon was described as the 'curse of dimensionality' by Bellman [1961]. The solution is that data dimensionality should be kept as low as possible consistent with the need to describe the particular research project adequately. Dimensionality reduction is the process of reducing the number of redundant variables under consideration, and can be divided into two major types: variable selection and variable extraction.

i. *Variable selection methods*

Variable selection methods try to identify a subset of the more important user-defined variables and to remove the remainder from the analysis (given some definition of importance) without losing too much information, thereby achieving dimensionality reduction. Given that variable selection methods aim to select a subset of the more important variables, a well-defined criterion of importance is fundamental. Two of the most often used ones in the literature are variable selection based on frequency and variable selection based on variance, and these are briefly described below. Others, such as variable selection based on term frequency-inverse document frequency (TF-IDF) and measures of nonrandomness, are also available, but these give results similar to those based on frequency and variance, and the additional complexity associated with them is therefore felt not to justify their inclusion; for further information on these see [e.g. Moisl, 2015; Belew, 2000; Salton & McGill, 1983; Robertson, 2004].

a. Variable selection based on frequency

Frequency is the simplest criterion for selecting features from a data matrix: those variables which occur most often in the research domain — in the present domain, words in text — are judged to be the most important, and lost which occur least often are taken to be least important and can therefore be discarded. With respect to clustering, the fundamental idea is that a variable should represent something which occurs often enough for it to make a significant contribution to the clustering of the data vectors. To select variables based on frequency, given an *m* x *n* frequency data matrix *D*; the value at *Dij* is the number of times variable *j*, for *j*=1…*n*, occurs in text *i*, for *i*=1…*m*. The frequency of occurrence of variable *j* across the entire corpus of texts is then:

$$freq(F_j) = \sum_{i=1..m} F_{i,j}$$

Frequencies of for all the columns data matrix *D* are calculated, sorted the variables in descending order of frequency, the most useful variables are selected and the less frequent variables are eliminated from *D*. Substantial dimensionality reduction can be achieved by applying this criterion to a data matrix *D*.

b. Variable selection based on variance

Variability refers to the amount of variation in the values that a variable takes. Any variable *x* is an interpretation of some aspect of the physical world, and a value assigned to *x* is a measurement of the world in terms of that interpretation. If *x* is to describe the ages of people, it can take different values for different persons or for the same person at different times. Unless all people are exactly the same age, or the age of the same person is fixed, the values which *x* takes will vary substantially, and can, therefore, contribute to the distinction of people from one another, or of the age of same person at different times (i.e. the more different people groups one tests, the more variation one will see in the ages). This possibility of variability in the values assigned to variable *x* gives it its descriptive utility: an identical value for *x* tells that what *x* stands for in the real world does not change, moderate variability in the value tells that aspect of the world changes only a little, and widely differing values tells that it changes substantially. In general, therefore, the possibility of variability in the values assigned to variables is necessary to the ability of variables to describe objects and thereby to represent reality. Clustering of texts or of anything else depends on there being variability in their characteristics; identical texts having the same stylistic descriptors cannot be meaningfully clustered. When the texts to be clustered are described by variables, then the variables are only useful for the purpose if there is significant variation in the values that they take. If, for example, a large number of people were described by their weights or heights, we would expect there to be logically substantial variation in values for each of them, and any cluster analysis method could legitimately be used to cluster them. On the other hand, if a large number of people were

© 2016 Global Journals Inc. (US)

described by variables like 'eyes', 'noses', and 'legs', there would be almost no or little variation or high correlation with other features, since, with very few exceptions, everyone has two eyes and a nose, and clustering based on these variables would be effectively useless. In any clustering application, therefore, one is looking for variables with substantial variation in their values, and can ignore variables with little or no variation. Variables with no or little variation should be removed from data matrix as they contain little information and complicate cluster analysis by making the data higher-dimensionality than it needs to be [Moisl, 2015].

Mathematically, the degree of variation in the values of a variable is described by its variance. The variance of a set of variable values is the average deviation of those values from their mean. Assume a set of n values $\{x_1, x_2...x_n\}$ assigned to a variable x. The mean of these values μ is $(x_1 + x_2 + ... + x_n)/n$. The amount by which any given value x_i differs from μ is then $x_i - \mu$. The mean difference from μ across all values is therefore $\Sigma_{i=1..n} (x_i - \mu)/n$. This mean difference of variable values from their mean almost but not quite corresponds to the definition of variance. One more step is necessary, and it is technical rather than conceptual. Because μ is an average, some of the variable values will be greater than μ, and some will be less. Consequently, some of the differences $(x_i - \mu)$ will be positive and some negative. When all the $(x_i - \mu)$ are added up, as above, they will cancel each other out. To prevent this, the $(x_i - \mu)$ are squared. The standard definition of variance for n values $\{x_1, x_2...x_n\}$ assigned to a variable x, therefore, is:

$$v = (\sum_{i=1..n} (x_i - \mu)^2) / n$$

To show how a variance is calculated, consider the following frequency counts of six variables (the, a, she, him, then, him) occurring in the corresponding five texts (a, b, c, d, e)

	the	a	she	him	then	he
Text.a	155	158	192	131	167	177
Text.b	43	70	76	64	58	69
Text.c	24	17	27	126	100	150
Text.d	73	89	100	190	50	60
Text.e	80	100	88	90	60	89

For text.a, the mean is 163.33, and the Std is:

$155-163.33=(-8.33)^2= 69.38$

$158-163.33=(-5.33)^2 =28.40$

$192-163.33=(29)^2 =841$

$131-163.33=(-32.33)^2= 1045$

$167-163.33=(3.67)^2 =13.46$

$177-163.33= (13.6)^2= 184.96$

$69.38+28.40+841+1045+13.46+184.96= 2182.2$ (the sum of squared of differences or standard deviations).

Thus the variance for text.a is $2182.2/6=363.7$

Doing the same calculation for the remaining texts, we have the following variances 100, 150, 190, 200 for texts b, c, d, and e respectively.

Given a data matrix M in which the row vectors are the texts and the column vectors are lexical type variables describing the texts, and also that the aim is to cluster analyze these texts on the basis of the differences among them, the application of variance/standard deviation to dimensionality reduction is straightforward: calculate and plot the variances of the columns and, if any have variability which is low in relation to that of the others, remove them on the grounds that they contribute little to differentiation of the texts, and decide on a threshold selection (the set of retained variables from each column of the data matrix).

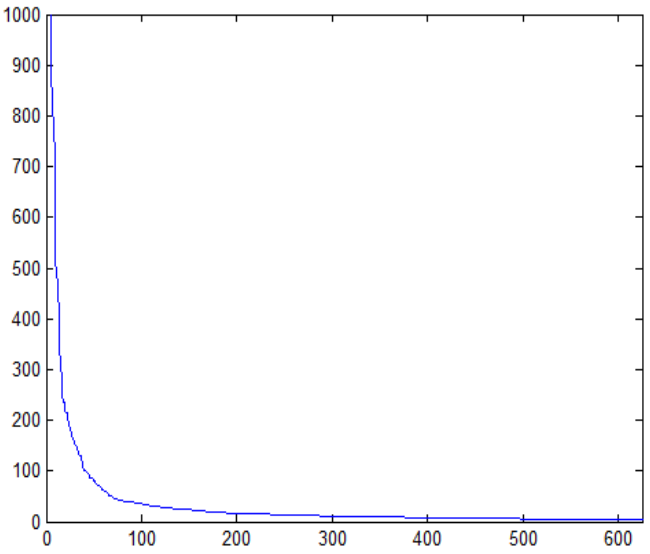


Figure 5 : Example of sorted variable variances after eliminating low-variance variables from the columns data matrix

Relative variance can now clearly be seen. The high-variance variables are on the left, and lower-variance ones on the right. The high-variance variables have to be kept, since they are the main criteria by which the NECTE speakers are distinguished. The flat area on the right represents the low-variance variables that contribute little or nothing to distinction among speakers, and these variables, starting about 175 and moving to the right, can be discarded.

For discussions that are concerned only with variable selection for clustering see, for example, [Dy, 2008; Dy and Bodley, 2004; Jain, Murty, and Flynn, 1999].

d) Variable extraction methods

Variable extraction methods replace the set of user-defined variables with a smaller set of variables which reduces dimensionality but captures most of the variability in the original set. These methods often

© 2016 Global Journals Inc. (US)

achieve a greater degree of dimensionality reduction, but at a cost: the newly-defined variables are generated by mathematical procedures, and their meaning relative to the research domain is typically difficult to determine reliably. There are a wide of variable extraction methods:

- Singular value decomposition (SVD)
- Principal Components Analysis (PCA)
- Factor Analysis (FA)
- Multi-dimensional Scaling (MDS)
- Isomap
- Self-Organizing Map (SOM)

Each one of these methods can be used for dimensionality reduction as a feature or variable extractor, and to visualize the clusters as a clustering method. The literature on these methods is extensive and this is just a brief outline that one can follow. A more comprehensive account can be found in, for example, [Moisl, 2015; Borg and Groenen, 2005; Kohonen, 2001; Tenenbaum, de Silva, and Langford, 2000; Gordon, 1999]. However, it will be useful to look briefly at one of these methods, that is, PCA, as a dimensionality reduction method, to see how it reduces the data down into basic components, removing any unnecessary variables.

Principal Components Analysis (PCA)is actually a dimensionality reduction method, which aims to transform a set of correlated variables into a -- usually smaller-- set of uncorrelated ones. PCA can also be used for clustering if the dimensionality is sufficiently reduced. The conceptual basis of PCA is elimination of variable redundancy. Specifically, given a matrix of m data items described by n variables, principal components analysis is a technique for redescribing the m items in terms of k variables, where $k < n$, such that most of the variability in the original n variables is retained. When $k = 2$ or $k = 3$ the m data items can be plotted in two or three dimensional space and any clusters can thereby be directly perceived. Relative to an n-dimensional data set D, the essence of PCA is this:

- An n-dimensional orthogonal basis for D is constructed, such that each axis is the least-squares best fit to one of the n directions of variation in D.
- The axes along which there is relatively little variation are eliminated, leaving an m-dimensional basis for D, where $m < n$.
- The original n-dimensional data D is projected into the reduced m-dimensional space, which yields a data set D' that is dimensionality-reduced but still contains most of the variability in D.

III. Conclusion

In this article, I discussed three techniques to adjust a data matrix before applying cluster analytical methods to take account of the variation in scales among the variables, the variation in length among the texts, and any superfluous variables in it using standardization, normalization, and dimensionality reduction techniques. A full and detailed consideration of each of these techniques addressed in this article would require several articles. My treatment of them is, necessarily, introductory and brief. Therefore, I urge interested computational linguists to follow the more in depth sources cited in the references. The application of these techniques for cluster analysis with specific reference to corpus linguistics is only one of many possibilities. The data items/matrix rows might be students in a second language learning (L2) survey and the variable/matrix columns motivational factors like learning experience, attitudes, cultural interest, and so on. n formants in a sociolinguistic or dialectological survey and the variables/matrix columns phonetic features like voicing, and so on. The lexical frequency example was selected because it is generic with respect to a wide range of possible applications.

IV. Acknowledgments

The author wishes to thank all those who dedicated their time answering my queries and providing me with valuable comments during the preparation of this study.

Conflicts of Interest

The author declares no conflict of interest.

References Références Referencias

1. Belew, R. (2000): Finding Out About: A Cognitive Perspective in Search Engine Technology and the WWW. Cambridge: Cambridge University Press.
2. Borg, I. and P. Groenen (2005): Modern Multi-dimensional Scaling. 2nd ed. Berlin: Springer.
3. Chu, C., J. Holliday, and P. Willett (2009): "Effect of data standardization on chemical clustering and similarity searching." In: Journal of Chemical Information and Modeling 49; 155–161.
4. Dy, J. (2008): "Unsupervised feature selection." In: Computational Methods of Feature Selection. Ed. by H. Liu and H. Motada. London: Chapman and Hall CRC; 19–39.
5. Dy, J. and C. Bodley (2004): "Feature selection for unsupervised learning."In: Journal of Machine Learning Research 5; 845–89.
6. Gnanandesikan, R., S. Tsao, and J. Kettenring (1995): "Weighting and selection of variables for cluster analysis." In: Journal of Classification 12; 113–136.
7. Gordon, A. (1999). Classification 2nd ed. London: Chapman and HallJain, A., M. Murty, and P. Flynn (1999): "Data clustering: a review." In: ACM Computing Surveys 31; 264–323.

© 2016 Global Journals Inc. (US)

8. Hermann Moisl. (2015). Cluster Analysis for Corpus Linguistics. Berlin: De Gruyter Mouton.
9. Kohonen, T. (2001). Self-Organizing Maps. 3rd ed. Berlin: Springer.
10. Milligan, G. andM. Cooper (1985): "An examination of procedures for determining the number of clusters in a data set." In: Psychometrika 50; 159–79.
11. Priddy, K. L. and Keller, P. E. (2005). Artificial Neural Networks: An Introduction. USA: Spie Press.
12. Singhal, A., C. Buckley, and M. Mitra (1996): "Pivoted document length normalization." In: Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR-96); 21–29.
13. Singhal, A. Salton, G., and Buckley, C. (1995): "Document Length Normalization." In: Information Processing and Management 32; 619–633.
14. Tenenbaum, J., V. deSilva, and J. Langford (2000): "A global geometric framework for nonlinear dimensionality reduction." In: Science 290; 2319–23.

© 2016 Global Journals Inc. (US)

Global Journals Inc. (US) Guidelines Handbook 2016

www.GlobalJournals.org

FELLOWS

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN HUMAN SCIENCE (FARSHS)

Global Journals Incorporate (USA) is accredited by Open Association of Research Society (OARS), U.S.A and in turn, awards "FARSHS" title to individuals. The 'FARSHS' title is accorded to a selected professional after the approval of the Editor-in-Chief/Editorial Board Members/Dean.

- The "FARSHS" is a dignified title which is accorded to a person's name viz. Dr. John E. Hall, Ph.D., FARSS or William Walldroff, M.S., FARSHS.

FARSHS accrediting is an honor. It authenticates your research activities. After recognition as FARSHS, you can add 'FARSHS' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, and Visiting Card etc.

The following benefits can be availed by you only for next three years from the date of certification:

FARSHS designated members are entitled to avail a 40% discount while publishing their research papers (of a single author) with Global Journals Incorporation (USA), if the same is accepted by Editorial Board/Peer Reviewers. If you are a main author or co-author in case of multiple authors, you will be entitled to avail discount of 10%.

Once FARSHS title is accorded, the Fellow is authorized to organize symposium/seminar/conference on behalf of Global Journal Incorporation (USA). The Fellow can also participate in conference/seminar/symposium organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent.

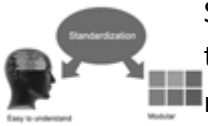
You may join as member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. In addition, it is also desirable that you should organize seminar/symposium/conference at least once.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time.This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

© Copyright by Global Journals Inc.(US) | Guidelines Handbook

The FARSHS can go through standards of OARS. You can also play vital role if you have any suggestions so that proper amendment can take place to improve the same for the benefit of entire research community.

As FARSHS, you will be given a renowned, secure and free professional email address with 100 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders,Auto-Responders, Email Delivery Route tracing, etc.



The FARSHS will be eligible for a free application of standardization of their researches. Standardization of research will be subject to acceptability within stipulated norms as the next step after publishing in a journal. We shall depute a team of specialized research professionals who will render their services for elevating your researches to next higher level, which is worldwide open standardization.

The FARSHS member can apply for grading and certification of standards of the educational and Institutional Degrees to Open Association of Research, Society U.S.A.
Once you are designated as FARSHS, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria. After certification of all your credentials by OARS, they will be published on your Fellow Profile link on website https://associationofresearch.org which will be helpful to upgrade the dignity.

The FARSHS members can avail the benefits of free research podcasting in Global Research Radio with their research documents. After publishing the work, (including published elsewhere worldwide with proper authorization) you can upload your research paper with your recorded voice or you can utilize chargeable services of our professional RJs to record your paper in their voice on request.

The FARSHS member also entitled to get the benefits of free research podcasting of their research documents through video clips. We can also streamline your conference videos and display your slides/ online slides and online research video clips at reasonable charges, on request.

© Copyright by Global Journals Inc.(US)| Guidelines Handbook

The FARSHS is eligible to earn from sales proceeds of his/her researches/reference/review Books or literature, while publishing with Global Journals. The FARSHS can decide whether he/she would like to publish his/her research in a closed manner. In this case, whenever readers purchase that individual research paper for reading, maximum 60% of its profit earned as royalty by Global Journals, will be credited to his/her bank account. The entire entitled amount will be credited to his/her bank account exceeding limit of minimum fixed balance. There is no minimum time limit for collection. The FARSS member can decide its price and we can help in making the right decision.

The FARSHS member is eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get remuneration of 15% of author fees, taken from the author of a respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN HUMAN SCIENCE (MARSHS)

The ' MARSHS ' title is accorded to a selected professional after the approval of the Editor-in-Chief / Editorial Board Members/Dean.

The "MARSHS" is a dignified ornament which is accorded to a person's name viz. Dr John E. Hall, Ph.D., MARSHS or William Walldroff, M.S., MARSHS.

MARSHS accrediting is an honor. It authenticates your research activities. Afterbecoming MARSHS, you can add 'MARSHS' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, Visiting Card and Name Plate etc.

The following benefitscan be availed by you only for next three years from the date of certification.

MARSHS designated members are entitled to avail a 25% discount while publishing their research papers (of a single author) in Global Journals Inc., if the same is accepted by our Editorial Board and Peer Reviewers. If you are a main author or co-author of a group of authors, you will get discount of 10%.

As MARSHS, you willbegiven a renowned, secure and free professional email address with 30 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders,Auto-Responders, Email Delivery Route tracing, etc.

© Copyright by Global Journals Inc.(US) | Guidelines Handbook

We shall provide you intimation regarding launching of e-version of journal of your stream time to time.This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The MARSHS member can apply for approval, grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A.

Once you are designated as MARSHS, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria.

It is mandatory to read all terms and conditions carefully.

© Copyright by Global Journals Inc.(US)| Guidelines Handbook

AUXILIARY MEMBERSHIPS

Institutional Fellow of Open Association of Research Society (USA) - OARS (USA)

Global Journals Incorporation (USA) is accredited by Open Association of Research Society, U.S.A (OARS) and in turn, affiliates research institutions as "Institutional Fellow of Open Association of Research Society" (IFOARS).

The "FARSC" is a dignified title which is accorded to a person's name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

The IFOARS institution is entitled to form a Board comprised of one Chairperson and three to five board members preferably from different streams. The Board will be recognized as "Institutional Board of Open Association of Research Society"-(IBOARS).

The Institute will be entitled to following benefits:

The IBOARS can initially review research papers of their institute and recommend them to publish with respective journal of Global Journals. It can also review the papers of other institutions after obtaining our consent. The second review will be done by peer reviewer of Global Journals Incorporation (USA) The Board is at liberty to appoint a peer reviewer with the approval of chairperson after consulting us.

The author fees of such paper may be waived off up to 40%.

The Global Journals Incorporation (USA) at its discretion can also refer double blind peer reviewed paper at their end to the board for the verification and to get recommendation for final stage of acceptance of publication.

The IBOARS can organize symposium/seminar/conference in their country on behalf of Global Journals Incorporation (USA)-OARS (USA). The terms and conditions can be discussed separately.

The Board can also play vital role by exploring and giving valuable suggestions regarding the Standards of "Open Association of Research Society, U.S.A (OARS)" so that proper amendment can take place for the benefit of entire research community. We shall provide details of particular standard only on receipt of request from the Board.

The board members can also join us as Individual Fellow with 40% discount on total fees applicable to Individual Fellow. They will be entitled to avail all the benefits as declared. Please visit Individual Fellow-sub menu of GlobalJournals.org to have more relevant details.

© Copyright by Global Journals Inc.(US) | Guidelines Handbook

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

After nomination of your institution as "Institutional Fellow" and constantly functioning successfully for one year, we can consider giving recognition to your institute to function as Regional/Zonal office on our behalf.
The board can also take up the additional allied activities for betterment after our consultation.

The following entitlements are applicable to individual Fellows:

Open Association of Research Society, U.S.A (OARS) By-laws states that an individual Fellow may use the designations as applicable, or the corresponding initials. The Credentials of individual Fellow and Associate designations signify that the individual has gained knowledge of the fundamental concepts. One is magnanimous and proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice.

Open Association of Research Society (US)/ Global Journals Incorporation (USA), as described in Corporate Statements, are educational, research publishing and professional membership organizations. Achieving our individual Fellow or Associate status is based mainly on meeting stated educational research requirements.

Disbursement of 40% Royalty earned through Global Journals : Researcher = 50%, Peer Reviewer = 37.50%, Institution = 12.50% E.g. Out of 40%, the 20% benefit should be passed on to researcher, 15 % benefit towards remuneration should be given to a reviewer and remaining 5% is to be retained by the institution.

We shall provide print version of 12 issues of any three journals [as per your requirement] out of our 38 journals worth $ 2376 USD.

Other:

The individual Fellow and Associate designations accredited by Open Association of Research Society (US) credentials signify guarantees following achievements:

- The professional accredited with Fellow honor, is entitled to various benefits viz. name, fame, honor, regular flow of income, secured bright future, social status etc.

- In addition to above, if one is single author, then entitled to 40% discount on publishing research paper and can get 10%discount if one is co-author or main author among group of authors.
- The Fellow can organize symposium/seminar/conference on behalf of Global Journals Incorporation (USA) and he/she can also attend the same organized by other institutes on behalf of Global Journals.
- The Fellow can become member of Editorial Board Member after completing 3yrs.
- The Fellow can earn 60% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.
- Fellow can also join as paid peer reviewer and earn 15% remuneration of author charges and can also get an opportunity to join as member of the Editorial Board of Global Journals Incorporation (USA)
- • This individual has learned the basic methods of applying those concepts and techniques to common challenging situations. This individual has further demonstrated an in–depth understanding of the application of suitable techniques to a particular area of research practice.

Note :

"

- In future, if the board feels the necessity to change any board member, the same can be done with the consent of the chairperson along with anyone board member without our approval.
- In case, the chairperson needs to be replaced then consent of 2/3rd board members are required and they are also required to jointly pass the resolution copy of which should be sent to us. In such case, it will be compulsory to obtain our approval before replacement.
- In case of "Difference of Opinion [if any]" among the Board members, our decision will be final and binding to everyone.

"

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.Online Submission: There are three ways to submit your paper:

(A) (I) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer, then Direct Submission through Homepage is also available.

(C) If these two are not conveninet , and then email the paper directly to dean@globaljournals.org.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Swis 721 Lt BT.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, "Abstract" word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be three lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.
Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals Inc. (US) welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals Inc. (US) is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global

Journals Inc. (US) are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals Inc. (US) and Editorial Board, will become the copyright of the Global Journals Inc. (US).

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals Inc. (US) follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads:

Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications

Research letters: The letters are small and concise comments on previously published matters.

5.STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also.Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

Before start writing a good quality Computer Science Research Paper, let us first understand what is Computer Science Research Paper? So, Computer Science Research Paper is the paper which is written by professionals or scientists who are associated to Computer Science and Information Technology, or doing research study in these areas. If you are novel to this field then you can consult about this field from your supervisor or guide.

TECHNIQUES FOR WRITING A GOOD QUALITY RESEARCH PAPER:

1. Choosing the topic: In most cases, the topic is searched by the interest of author but it can be also suggested by the guides. You can have several topics and then you can judge that in which topic or subject you are finding yourself most comfortable. This can be done by asking several questions to yourself, like Will I be able to carry our search in this area? Will I find all necessary recourses to accomplish the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. Pick a good study spot: To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

© Copyright by Global Journals Inc.(US)| Guidelines Handbook

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. Think technically: Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page
- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

© Copyright by Global Journals Inc.(US)| Guidelines Handbook

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.

- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

© Copyright by Global Journals Inc.(US)| Guidelines Handbook

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.
- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.

- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

© Copyright by Global Journals Inc.(US)| Guidelines Handbook

Criterion for Grading a Research Paper (Compilation) by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

INDEX

E

Elliptical · 46

G

Generosity · 19

I

Immunized · 2
Inception · 6, 8, 9
Intertwined · 26
Intonation · 53
Intuitive · 24

L

Lexical · 60, 61, 62, 63, 64

M

Morphemes · 13

S

Sparse · 62
Stereotyped · 3
Strides · 5, 9

T

Tidings · 14

save our planet

Global Journal of Human Social Science

Visit us on the Web at www.GlobalJournals.org | www.SocialScienceResearch.org
or email us at helpdesk@globaljournals.org

ISSN 975587

© Global Journals