Taraka Rama

Studies in computational historical linguistics

Data linguistica

<http://www.svenska.gu.se/publikationer/data-linguistica/>

Editor: Lars Borin

Språkbanken
Department of Swedish
University of Gothenburg

27 • 2015

Taraka Rama

Studies in computational historical linguistics

Models and analyses

Gothenburg 2015

Data linguistica 27
ISBN 978-91-87850-58-5
ISSN 0347-948X
E-publication http://hdl.handle.net/2077/40571

Printed in Sweden by
Taberg Media Group AB 2015

Typeset in LaTeX 2ε by the author

Cover design by Kjell Edgren, Informat.se, and Sven Lindström

Front cover illustration:
Maps of the relationships between Indo-European and Uralic languages by *Minna Sundberg*.

Author photo on back cover by Kristina Holmlid

ABSTRACT

Computational analysis of historical and typological data has made great progress in the last fifteen years. In this thesis, I work with vocabulary lists for addressing some classical problems in historical linguistics such as cognate identification, discriminating related languages from unrelated languages, assigning possible dates to splits in a language family, and providing an internal structure to a language family. I compare the internal structure inferred from vocabulary lists with the family trees given in Ethnologue. I explore the ranking of lexical items in the widely used Swadesh word list and compare my ranking to another quantitative reranking method and short word lists composed for discovering long-distance genetic relationships. I show that the choice of string similarity measures is important for internal classification and for discriminating related from unrelated languages. The dating system presented in this thesis can be used for assigning age estimates to any new language group and overcomes the assumption of a constant rate of lexical replacement assumed by glottochronology. I train and test a linear classifier based on gap-weighted subsequence features for the purpose of cognate identification. An important conclusion from these results is that n-gram approaches can be used for different historical linguistic purposes.

SAMMANFATTNING

Datorbaserad forskning på språkhistoriska och språktypologiska data har tagit stora steg framåt under den senaste femtonårsperioden. I denna avhandling presenterar jag mitt arbete där jag använder mig av storskalig datorbearbetning av listor med ett språks grundläggande ordförråd (så kallade Swadeshlistor) för en stor mängd språk i syfte att med ny metodologi belysa ett antal klassiska problem inom den historiska språkvetenskapen. Exempel på sådana problem är: att identifiera besläktade ord (kognater) mellan språk, att skilja besläktade och obesläktade språk åt, att uppskatta tidpunkter för språksplittring (då ett antaget urspråk har delats upp i två eller flera dotterspråk), samt att fastställa de interna släktskapsförhållandena i språkfamiljer. Jag jämför de språkfamiljsstrukturer jag får fram genom datorbearbetning av ordlistor med de familjeträd som anges i standardreferensen Ethnologue. Jag undersöker vokabulären i olika Swadeshlistor och även andra föreslagna liknande ordlistor med avseende på deras användbarhet för att studera ovannämnda problem. Jag visar att valet av strängjämförelsemetod är centralt för att skilja besläktade och obesläktade språk åt och för att fastställa en språkfamiljs interna struktur. Den dateringsmetod som utvecklas och presenteras i denna avhandling kan användas för att uppskatta tidpunkter för språksplittring i vilken språkfamilj som helst. En viktig egenskap hos metoden är att den inte är baserad på glottokronologins grundantagande om att enheterna i ett språks basordförråd byts ut i konstant takt. I avhandlingen utvecklar och utvärderar jag en metod för kognatidentifiering som bygger på maskininlärning med användning av särdrag baserade på en transformerad representation av vokabulärdata, i form av en viss sorts diskontinuerliga delsekvenser. En viktig slutsats av de resultat som presenteras i denna avhandling är att n-gram-baserade metoder lämpar sig väl för att angripa ett antal språkhistoriska forskningsfrågor.

ACKNOWLEDGEMENTS

I thank my supervisors, family, friends, collaborators, teachers, opponents, reviewers, and colleagues for everything. This thesis would not have been possible without them. The paper presentations would not have been possible without the generous funding of the Graduate School in Language Technology, the Centre for Language Technology, and the Digital Areal Linguistics project.

CONTENTS

Appendices

1 INTRODUCTION

This thesis can be viewed as an attempt at applying techniques from *Language Technology* (LT; also known as Natural Language Processing [NLP] or Computational Linguistics [CL]) to the traditional historical linguistics problems such as identification of cognates, dating of language families, and language classification.

Modern humans appeared on this planet about 100,000–150,000 years ago (Vigilant et al. 1991; Nettle 1999a). Given that all modern humans descended from a small African ancestral population, did all the 7,000 languages (Lewis, Simons and Fennig 2013) descend from a common language? Did language emerge from a single source (*monogenesis*) or from multiple sources at different times (*polygenesis*)? Do these languages fall under a single family – descended from a single language which is no longer spoken – or multiple families? If they fall under multiple families, how are they related to each other? What is the internal structure of a single language family? How old is a family or how old are the intermediary members of a family? Can we give reliable age estimates to these languages? This thesis attempts to answer some of these questions.

1.1 Computational historical linguistics

This section gives a brief introduction to historical linguistics and then to the related field of computational historical linguistics.[1]

1.1.1 Historical linguistics

Historical linguistics is the oldest branch of modern linguistics. Historical linguistics is concerned with language change, the processes introducing the language change and also identifying the (pre-)historic relationships between

[1] To the best of my knowledge, Lowe and Mazaudon (1994) were the first to use the term.

languages (Trask 2000: 150). Historical linguistics works towards identifying the not-so-apparent relations between languages. Historical linguistics has succeeded in identifying the relation between languages spoken in the Indian sub-continent, the Uyghur region of China, and Europe (the Indo-European family); as well as between languages spoken in Madagascar islands and the remote islands in the Pacific Ocean (the Austronesian family).

A subbranch of historical linguistics is comparative linguistics. According to Trask (2000: 65), comparative linguistics is a branch of historical linguistics which seeks to identify and elucidate genetic relationships among languages. Comparative linguistics works through the comparison of *linguistic systems*. The comparative linguists compare vocabulary items (not any but following a few general guidelines) and morphological forms; and accumulate the evidence for language change through systematic sound correspondences (and sound shifts) to propose connections between languages descended through modification from a common ancestor.

The work reported in this thesis is related to the application of computational techniques to address the traditional problems in historical linguistics.

1.1.2 Computational historical linguistics

Computational historical linguistics (CHL) aims to design computational methods to identify linguistic differences between languages based on different aspects of language: phonology, morphology, lexicon, and syntax. CHL also includes computational simulations of language change in speech communities (Nettle 1999b), simulation of disintegration (divergence) of proto-languages (De Oliveira, Sousa and Wichmann 2013), the relation between population sizes and rate of language change (Wichmann and Holman 2009a), and simulation of the current distribution of language families (De Oliveira et al. 2008). Finally, CHL proposes and studies formal and computational models of linguistic evolution through language acquisition (Briscoe 2002) and computational and evolutionary aspects of language (Nowak, Komarova and Niyogi 2002; Niyogi 2006).

The use of mathematical and statistical techniques to classify languages (Kroeber and Chrétien 1937) and evaluate language relatedness hypotheses (Kroeber and Chrétien 1939; Ross 1950; Ellegård 1959) has been attempted in the past. Swadesh (1950) invented the method of lexicostatistics which works with standardized vocabulary lists but the similarity judgment between the words is based on cognacy rather than the superficial word form similarity technique of multilateral comparison (Greenberg 1993: cf. section 2.4.2).

Swadesh (1950) uses *cognate* counts to posit internal relationships in a subgroup of a language family. Cognates are related words across languages whose origin can be traced back to a (reconstructed or documented) word in a common ancestor. Cognates are words such as Sanskrit *dva* and Armenian *erku* 'two' whose origin can be traced back to a common ancestor. Cognates usually have similar form and also similar meaning and are not borrowings (Hock 1991: 583–584). Traditionally cognates were not identified through a computer but by a manual procedure.

Hewson 1973 (see Hewson 2010 for a more recent description) can be considered the first study where computers were used to reconstruct the words of a proto-language – specifically Proto-Algonquian (the common ancestor of the Algonquian language family). Dictionaries of four Algonquian languages – Fox, Cree, Ojibwa, and Menominee – were converted into computer-readable format – skeletal forms, only the consonants are fed into the computer and vowels are omitted – and then an ancestral form for a word form was projected by searching through all possible sound-correspondences. The projected proto-forms for each language were alphabetically sorted to yield a set of putative proto-forms for the four languages. Finally, a linguist with sufficient knowledge of the language family would then go through the putative proto-list and remove the unfeasible cognates.

In practice, historical linguists mostly work with word lists – selected words which are not nursery forms, onomatopoeic forms, chance similarities, and borrowings (Campbell 2003). Dictionaries are a natural extension to word lists (Wilks, Slator and Guthrie 1996). Assuming that we are provided with bilingual dictionaries of some languages, can we simulate the task of a historical linguist? How far can we automate the steps of weeding out borrowings, extracting sound correspondences, and positing relationships between languages? An orthogonal task to language comparison is the task of comparing the earlier forms of an extant language to its modern form.

A related task in comparative linguistics is internal reconstruction. Internal reconstruction seeks to identify the exceptions to patterns present in extant languages and then reconstruct the regular patterns in the older stages. The laryngeal hypothesis for Proto-Indo-European (PIE) is a classical case of internal reconstruction. Saussure (1879) applied internal reconstruction to explain the aberrations in the reconstructed root structures of PIE.

An automatic mapping of the words in digitized text from the middle ages to the current forms would be a CHL task. Another task would be to identify variations in written forms and normalize the orthographic variations. These tasks fall within the field of *NLP for historical texts* (Piotrowski 2012). For instance, deriving the suppletive verbs such as *go, went* or adjectives *good, better, best* from ancestral forms or automatically identifying the

corresponding cognates in Sanskrit would also be a CHL task.

There has been a renewed interest in the application of computational and quantitative techniques to problems in historical linguistics for the last fifteen years. This new wave of publications has been met with initial skepticism which lingers from the past of glottochronology.[2] However, the initial skepticism has given way to consistent work in terms of methods (Agarwal and Adams 2007), workshop(s) (Nerbonne and Hinrichs 2006), journals (Wichmann and Good 2011), and an edited volume (Borin and Saxena 2013).

The new wave of CHL publications are co-authored by linguists, computer scientists, computational linguists, physicists, and evolutionary biologists. Except for some scattered efforts (Kay 1964; Sankoff 1969; Klein, Kuppin and Meives 1969; Durham and Rogers 1969; Smith 1969; Wang 1969; Dobson et al. 1972; Embleton 1986; Borin 1988; Dyen, Kruskal and Black 1992; Kessler 1995; Warnow 1997; Huffman 1998; Nerbonne, Heeringa and Kleiweg 1999), the area was not very active until the work of Gray and Jordan 2000, Ringe, Warnow and Taylor 2002, and Gray and Atkinson 2003. Gray and Atkinson (2003) employed Bayesian inference techniques, originally developed in computational biology for inferring the family trees of species, to infer the family tree of the Indo-European family based on the lexical cognate data. In LT, Bouchard-Côté et al. (2013) employed Bayesian inference techniques to reconstruct Proto-Austronesian forms for fixed-length word lists belonging to more than 400 modern Austronesian languages.

The work reported in this thesis is related to the well-studied problems of approximate matching of string queries in database records using string similarity measures (Gravano et al. 2001), automatic identification of languages in a multilingual text through the use of character n-grams and skip-grams, approximate string matching for cross-lingual information retrieval (Järvelin, Järvelin and Järvelin 2007), and ranking of documents in a document retrieval task. The description of the tasks and the motivation and its relation to the work reported in the thesis are given below.

The task of approximate string matching of queries with database records is related to the task of cognate identification. As noted before, another related but sort of inverse task is the detection of borrowings. Lexical borrowings are words borrowed into a language from an external source. Undetected lexical borrowings can lead to spurious affiliation between languages under consideration. For instance, English borrowed a lot of words from the Indo-Aryan languages (Yule and Burnell 1996) such as *bungalow, chutney, shampoo,* and *yoga*. If we base a genetic comparison on these borrowed words, the comparison would suggest that English is more closely

[2] See Nichols and Warnow (2008) for a survey on this topic.

related to the Indo-Aryan languages than to the other languages of IE family. Sometimes words can look similar due to random chance. Such words are known as *chance similarities*. One example from Bloomfield 1935 is Modern Greek *mati* and Malay *mata* 'eye'. However, these Modern Greek and Malay are unrelated and the words are similar only by coincidence.

The task of automated language identification (Cavnar and Trenkle 1994) can be related to the task of automated language classification. A typical language identifier system consists of multilingual character n-gram models, where each character n-gram model corresponds to a single language. A character n-gram model is trained on a set of texts of a language. The test set consisting of a multilingual text is matched to each of these language models to yield a probable list of languages to which each word in the test set belongs.

Relating to the automated language classification, an n-gram model can be trained on a word list for each language and all pair-wise comparisons of the n-gram models would yield a matrix of (dis)similarities – depending on the choice of similarity/distance measure – between the languages. These pair-wise matrix scores are supplied as an input to a clustering algorithm to infer a hierarchical structure of the languages.

Until now, I have listed and related the parallels between various challenges faced by a traditional historical linguist and the challenges in CHL. LT methods are employed to address research questions within the computational historical linguistics field. Examples of such applications are listed below.

- *Historical word form analysis*. Applying string similarity measures to map orthographically variant word forms in Old Swedish to the lemmas in an Old Swedish dictionary (Adesam, Ahlberg and Bouma 2012).
- *Deciphering extinct scripts*. Character n-grams (along with symbol entropy) have been employed to decipher foreign languages (Ravi and Knight 2008). Reddy and Knight (2011) analyze an undeciphered manuscript using character n-grams.
- *Tracking language change*. Tracking semantic change (Gulordava and Baroni 2011), [3] orthographic changes, and grammaticalization over time through the analysis of corpora (Borin et al. 2013).
- *Creation of language resources for older language from current day language*. SMT (Statistical Machine Translation) techniques are applied to annotate historical corpora of 14th century Icelandic through current-day Icelandic (Pettersson, Megyesi and Tiedemann 2013).

[3] How lexical items acquire a different meaning and function over time.

1.2 Summary and contributions

The following problems in historical linguistics are addressed through the application of computational techniques from LT:

1.2.1 Summary of the publications

I. *Phonological diversity and distance from Africa.* In publication I, we address the claim of Atkinson 2011 that phoneme diversity in the world's languages decreases as one moves away from Africa. Atkinson finds a negative correlation between the phoneme inventory size and distance from Africa in a dataset of 500 languages. Atkinson also finds that languages with large population sizes tend to have larger phoneme inventories. In publication I, we test these claims on a large database of more than 3000 languages and find that there is some support to Atkinson's claim. We also find that languages with large phoneme inventories tend to have shorter words and languages with larger speaker populations tend to have larger phoneme inventory sizes. The publication is co-authored with Søren Wichmann and Eric W. Holman and published in the peer-reviewed journal *Linguistic Typology* (Wichmann, Rama and Holman 2011).

II. *Lexical Item stability.* The task here is to generate a ranked list of concepts which can be used for investigating the problem of automatic language classification. Publication II, titled *N-gram approaches to the historical dynamics of basic vocabulary*, presents the results of the application of n-gram techniques to the vocabulary lists for 190 languages. In this work, we apply *n-gram (language models)* – widely used in LT tasks such as statistical machine translation (SMT), automated language identification, and automated drug detection (Kondrak and Dorr 2006) – to determine which concepts are resistant to the effects of time and geography. The results suggest that the ranked item list largely agrees with two other vocabulary lists proposed for identifying long-distance relationship. The paper is co-authored with Lars Borin and is published in the peer-reviewed *Journal of Quantitative Linguistics* (Rama and Borin 2013).

III. *Structural similarity and genetic classification.* How well can structural relations be employed for the task of language classification? In the publication III, titled *How good are typological distances for determining genealogical relationships among languages?*, we apply

different vector similarity measures to typological data for the task of language classification. We apply 14 vector similarity techniques, originally developed in the field of IE/IR, for computing the structural similarity between languages. The paper is co-authored with Prasanth Kolachina and is published as a short paper in the proceedings of *COLING 2012* (Rama and Kolachina 2012).

IV. *Estimating age of language groups.* In this task, we develop a system for dating the split/divergence of language groups present in the world's language families. Quantitative dating of language splits is typically associated with glottochronology (a criticized quantitative technique which assumes that the rate of lexical replacement for a time unit [1000 years] in a language is constant; Atkinson and Gray 2006). Publication IV, titled *Phonotactic diversity and time depth of language families*, presents a n-gram based method for automatic dating of the world's languages. We apply n-gram techniques to a carefully selected set of languages from different language families to yield baseline dates. This work is solely authored by me and is published in the peer-reviewed open source journal *PloS ONE* (Rama 2013).

V. *Genetic vs. areal linguistics in South Asia.* In publication V, we work with word lists – digitized from the Linguistic Survey of India (Grierson 1927) which predates Swadesh 1950 – of languages belonging to four different families that are spoken in South Asia. In the paper, we attempt to see if lexical comparison yields genetic classification or reflects the language contact situation reported in South Asia. The edit distance based language distance measure employed in the paper clusters languages into genetic groups rather than geographical proximal groups. This paper is co-authored with Lars Borin, Anju Saxena, and Bernard Comrie and is published in the peer-reviewed proceedings of the *Ninth International Conference on Language Resources and Evaluation* (Borin et al. 2014).

VI. *Comparison of string similarity measures for automated language classification.* A researcher attempting to carry out an automatic language classification for a set of languages is confronted with the following methodological problem. Which string similarity measure is the best for the tasks of discriminating related languages from the rest of unrelated languages and also for the task of determining the historical configuration of the related languages? Publication VI, *Evaluation of similarity measures for automatic language classification* is a book chapter that discusses the application of 14 string similarity measures to a dataset constituting more than half of the world's languages. In this

paper, we apply a statistical significance testing procedure to rank the performance of string similarity measures based on pair-wise similarity measures. This paper is co-authored with Lars Borin and is published in an edited volume, *Sequences in Language and Text* (Rama and Borin 2015).

VII. *Gap-weighted subsequences for cognate identification.* In the publication, I test if gap-weighted subsequences perform better than other string similarity measures for the purpose of cognate identification. In the paper, I find that a linear classifier trained on gap-weighted subsequences performs better than one trained on string similarity measures. The linear classifier is trained and tested on the Indo-European data of Dyen, Kruskal and Black 1992. The publication is solely authored by me and is published as a short paper in the proceedings of *NAACL 2015*.

1.2.2 Contributions

The contributions of the thesis are summarized below:

- In publication I, we establish that the number of unique phoneme segments in a short, standardized word list can be used as a proxy for the number of phoneme segments in a language. The paper also reports the following findings. There is an inverse correlation between word length and phoneme inventory sizes. Languages with large population sizes are associated with large phoneme inventory sizes and phoneme inventory sizes tend to decrease with a language's distance from Africa.

- In publication II, we develop an n-gram based procedure for ranking the items in a vocabulary list. The paper uses 100-word Swadesh lists as the point of departure and works with more than 150 languages. The n-gram based procedure shows that n-grams, in various guises, can be used for quantifying the resistance to lexical replacement across the branches of a language family.

- In publication III, we attempt to address the following three tasks: (i) Comparison of vector similarity measures for computing typological distances; (ii) correlating typological distances with genealogical classification derived from historical linguistics; (iii) correlating typological distances with the lexical distances computed from 40-word Swadesh lists. The paper also employs visualizations to show the strength and direction of correlations.

- In publication IV, I introduce phonotactic diversity as a measure of language divergence, language group size, and age of language groups. The combination of phonotactic diversity and lexical divergence are used to predict the dates of splits for more than 50 language families.

- In publication V, we test if the automatic lexical comparison of word lists from South Asia yield genetic or areal clusters. In this paper, we find that lexical comparison does not show contact situation but groups the languages into their respective families. In fact, the pair-wise lexical distances within the family correlate strongly with the family internal classification given in Ethnologue (Lewis, Simons and Fennig 2013).

- It has been observed that a particular string distance measure (Levenshtein distance or its phonetic variants: McMahon et al. 2007; Huff and Lonsdale 2011) is used for language distance computation purposes. However, string similarities is a well researched topic in computer science (Smyth 2003) and computer scientists have developed various string similarity measures for many practical applications. There is certainly a gap in CHL literature regarding the performance of other string similarity measures at the tasks of automatic language classification and inference of internal structures of language families. We attempt to fill this gap through publication VI. The paper compares the performance of 14 different string similarity techniques for the aforementioned purpose and applies a multiple-testing procedure for ranking the string similarity measures.

- In publication VII, I perform explicit cognate identification on the Indo-European database of Dyen, Kruskal and Black 1992. The paper employs gap-weighted subsequences as features for training a linear classifier to classify if two words are cognates or not. The paper is the first to employ the idea of gap-weighted subsequences from *String Kernels* (Lodhi et al. 2002) to capture the similarity between two words.

1.3 Overview of the thesis

The thesis is organized as follows. The first part of the thesis provides a background to the topic of computational historical linguistics. The second part of the thesis provides a linguistic discussion of the results reported in the third part.

Chapter 2 introduces the background in historical linguistics and discusses the different methods used in this thesis from a linguistic perspective. In this chapter, the concepts of sound change, semantic change, structural change, reconstruction, language family, core vocabulary, time-depth of language families, item stability, models of language change, and automated language classification are introduced and discussed. This chapter also discusses the comparative method in relation to the learning paradigm of semi-supervised learning that is widely used in LT (Yarowsky 1995; Abney 2004, 2010). Subsequently, the chapter proceeds to discuss the related computational work in the domain of automated language classification.

Chapter 3 discusses different linguistic databases developed during the last fifteen years. Although each chapter in part III has a section on linguistic databases, the motivation for the databases' development is not considered in detail in each paper.

Chapters 4 and *5* discuss the results of the publications reproduced in part III from a linguistic point of view. Chapter 4 discusses the relation between string similarity and cognates and motivates how string similarity measures capture some aspects of sound change. Chapter 5 discusses how phonotactic diversity is employed to predict ages for different language groups. The chapter explains the Gamma regression model and how it captures variation and rate of language change in different families.

Chapter 6 summarizes and concludes the background to the thesis and discusses future work.

Part III of the thesis consists of seven peer-reviewed publications, reformatted to fit the stylesheet of this series, but where the text is in all cases that of the original publication, leading to some repetition. Publication I was co-authored with Søren Wichmann and Eric W. Holman. Myself and Søren Wichmann performed the experiments. All the authors participated in the design of the experiments and the discussion of the results. The paper was written by Søren Wichmann and hence his name is listed as the top author. All the experiments in the publications III, IV, V, VI, and VII were conducted by me. The experiments in publication II were designed and conducted by myself and Prasanth Kolachina. The paper was written by myself and Prasanth Kolachina. In publications III and VI analysis of the results and the writing of the paper were performed by myself and Lars Borin. All the co-authors contributed to the writing of paper V. The experiments in publication IV and VII were designed and performed by myself. I am the sole author of publication IV and VII.

The following publications are not included in the thesis but were published or are under review during the last five years:

1. Kolachina, Sudheer, Taraka Rama and B. Lakshmi Bai 2011. Maximum parsimony method in the subgrouping of Dravidian languages. *Quantitative Investigations in Theoretical Linguistics* 4: 52–56.

2. Rama, Taraka and Lars Borin 2011. Estimating Language Relationships from a Parallel Corpus. A Study of the Europarl Corpus. *NEALT Proceedings Series (NODALIDA 2011 Conference Proceedings)*: 161–167.

3. Wichmann, Søren, Eric W. Holman, Taraka Rama and Robert S. Walker 2011. Correlates of reticulation in linguistic phylogenies. *Language Dynamics and Change* 1 (2): 205–240.

4. Rama, Taraka and Sudheer Kolachina 2013. Distance-based phylogenetic inference algorithms in the subgrouping of Dravidian languages. Lars Borin and Anju Saxena (eds), *Approaches to measuring linguistic differences*, 141–174. Berlin: De Gruyter, Mouton.

5. Rama, Taraka, Prasant Kolachina and Sudheer Kolachina 2013. Two methods for automatic identification of cognates. *Quantitative Investigations in Theoretical Linguistics* 5: 76.

6. Wichmann, Søren and Taraka Rama. Submitted. Jackknifing the black sheep: ASJP classification performance and Austronesian. For the proceedings of the symposium "Let's talk about trees", National Museum of Ethnology, Osaka, Febr. 9-10, 2013.

Part I

Background

2 COMPUTATIONAL HISTORICAL LINGUISTICS

This chapter is a survey of the terminology used in the publications reproduced in part III of the thesis. It covers related work on the topics of linguistic diversity, processes of language change, computational modeling of language change, units of genealogical classification, core vocabulary, time-depth, automated language classification, item stability, and corpus-based historical linguistics.

2.1 Differences and diversity

As noted in chapter 1, there are more than 7,000 living languages in the world according to the *Ethnologue* (Lewis, Simons and Fennig 2013) falling into more than 400 families (Nordhoff and Hammarström 2012). Many more languages went extinct during the last five hundred years due to war, disease, and language shift. It is believed that many more languages were spoken which did not leave any trace before they went extinct. The following questions are relevant with respect to linguistic differences and diversity:

- How different are languages from each other?
- Given that there are multiple families of languages, what is the variation inside each family? How divergent are the languages falling in the same family?
- What are the shared and differing linguistic aspects in a language family?
- How do we arrive at a numerical estimate of the differences? What are the units of such comparison?
- How and why do these differences arise?

The above questions are addressed in the recent framework of evolutionary linguistics (Croft 2000), which attempt to explain the language differences in the evolutionary biology frameworks of Dawkins 2006 and Hull 2001. Darwin (1871) himself had noted the parallels between biological evolution and language evolution. Atkinson and Gray (2005) provide a historical survey of the parallels between biology and language. Darwin makes the following statement regarding the parallels (Darwin 1871: 89–90):

> The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel [...] We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation.

Atkinson and Gray (2005) also observe that there has been a cross-pollination of ideas between biology and linguistics before Darwin. Table 2.1 summarizes the parallels between biological and linguistic evolution. I prefer to see the table as a guideline rather than a hard fact due to the following reasons:

- Biological drift is not the same as linguistic drift. Biological drift is random change in gene frequencies whereas linguistic drift is the tendency of a language to keep changing in the same direction over several generations (Trask 2000: 98).

- Ancient texts do not contain all the necessary information to assist a comparative linguist in drawing the language family history but a sufficient sample of DNA (extracted from a well-preserved fossil) can be compared to other biological family members to draw a family tree. For instance, the well-preserved finger bone of a species of *Homo* family (from the Denisova cave in Russia; known as Denisovan) was compared to Neanderthals and modern humans. The comparison showed that Neanderthals, modern humans, and Denisovans shared a common ancestor (Krause et al. 2010).

Croft (2008) summarizes the various efforts to explain the linguistic differences in the framework of evolutionary linguistics. Croft also notes that historical linguists have employed biological metaphors or analogies to explain language change; and then summarized the various evolutionary linguistic frameworks to explain language change. In evolutionary biology, some entity replicates itself either perfectly or imperfectly over time. The differences resulting from imperfect replication leads to differences in a

Biological evolution	Linguistic evolution
Discrete characters	Lexicon, syntax, and phonology
Homologies	Cognates
Mutation	Innovation
Drift	Drift
Natural selection	Social selection
Cladogenesis	Lineage splits
Horizontal gene transfer	Borrowing
Plant hybrids	Language Creoles
Correlated genotypes/phenotypes	Correlated cultural terms
Geographic clines	Dialects/dialect chains
Fossils	Ancient texts
Extinction	Language death

Table 2.1: Parallels between biological and linguistic evolution (Atkinson and Gray 2005).

population of species that over the time leads to splitting of the same species into different species. The evolutionary change is a two-step process:

- A variant is generated in the replication process.
- A variant is selected from the pool of variants.

Dawkins (2006) employs the selfish-gene concept according to which an organism is only a vector for the replication of the gene. The gene itself is generalized as a replicator. Dawkins and Hull differ slightly from each other with respect to selection of the variants. For Dawkins, the organism exists for replication whereas, for Hull, the selection is a function of the organism.

In linguistics, Ritt (2004) proposed a phonological change model which operates in the Dawkinsian framework. According to Ritt, phonemes, morphemes, phonotactic patterns, and phonological rules are replicators which are replicated through imitation. The process of imperfect imitation generates the variations in the linguistic behavior observed in a speech community. In this model, the linguistic utterance exists for the sake of replication rather than communication purposes.

Croft (2000, 2008) coins the term *lingueme* to denote a linguistic replicator. A lingueme is a token of linguistic structure produced in an utterance. A lingueme is a linguistic replicator and the interaction of the speakers (through production and comprehension) with each other causes the

generation and propagation of variation. Selection of particular variants is motivated through differential weighting of replicators in evolutionary biological models. The intentional and non-intentional mechanisms such as the pressure for mutual understanding and the pressure to conform to a standard variety cause imperfect replication in Croft's model. The speaker himself selects the variants fit for production, Nettle (1999a) however argues that functional pressure also operates in the selection of variants.

The cumulative differences induced through generations of imperfect replication cause linguistic diversity. Nettle (1999a: 10) lists the following types of linguistic diversity:

- *Language diversity*. This is the number of languages present in a given geographical area. New Guinea has the highest language diversity with more than 800 languages spoken in a relatively small island whereas Iceland has only one language (not counting immigration in recent history).

- *Phylogenetic diversity*. This is the number of (sub)families found in an area. For instance, India is rich in language diversity but has only four language families whereas South America has 108 language families (Campbell 2012: 67–69).

- *Structural diversity*. This is the number of languages found in an area with respect to a particular linguistic parameter. A linguistic parameter can be word order, size of phoneme inventory, morphological type, or suffixing vs. prefixing.

A different measure of diversity or differences is based on phonological similarities between two languages. Lohr (1998: chapter 3) introduces phonological methods for the genetic classification of European languages. The similarity between the phonetic inventories of individual languages is taken as a measure of language relatedness. Lohr (1998) also compares the same languages based on phonotactic similarity to infer a *phenetic* tree for the languages. It has to be noted that Lohr's comparison is based on hand-picked phonotactic constraints rather than constraints that are extracted automatically from corpora or dictionaries. Rama (2013) introduces phonotactic diversity as an index of age of language group and family size.

In an introduction to the volume titled *Approaches to measuring linguistic differences*, Borin (2013: 4) observes that we need to fix the units of comparison before attempting to measure the differences between the units. In the field of historical linguistics, language is the unit of comparison. In the closely related field of dialectology, dialectologists work with much thinner

samples of a single language. Namely, they work with language varieties (dialects) spoken at different sites in the geographical area where the language is spoken.

2.2 Language change

Language changes in different aspects: phonology, morphology, syntax, meaning, lexicon, and structure. Historical linguists gather evidence of language change from all possible sources and then use the information to classify languages. Thus, it is very important to understand the different kinds of language change for computational modeling of language change. In this section, the different processes of language change are described through examples from the Indo-European and Dravidian language families. Each description of a type of language change is followed by a description of the computational modeling of the respective language change.

2.2.1 Sound change

Sound change is the most studied of all the language changes (Crowley and Bowern 2009: 184). The typology of sound changes described in the following subsections indicate that the sound changes depend on the notions of position in the word, neighboring sounds (context) and the quality of the sound in focus. The typology of the sound changes is followed by a subsection describing the various string similarity algorithms which model different sound changes and hence are employed to compute the distance between a pair of cognates, a proto-form and its reflexes.

2.2.1.1 Lenition and fortition

Lenition is a sound change causing a sound to become less consonant-like. Consonants can undergo a shift from left to right on one of the scales given below in Trask (1996: 56).

- geminate > simplex
- stop > fricative > approximant
- stop > liquid
- oral stop > glottal stop
- non-nasal > nasal
- voiceless > voiced

A few examples (from Trask 1996) involving the movement of sound according to the above scales is as follows. Latin *cuppa* 'cup' > Spanish *copa*. Rhotacism, /s/ > /r/, in Pre-Latin is an example of this change where the genitive form of 'flower' **flosis* > *floris*.[4] Latin *faba* 'bean' > Italian *fava* is an example of fricativization. Latin *strata* > Italian *strada* 'road' is an example of voicing. The opposite of lenition is fortition, where a sound moves from right to left on one of the above scales. Fortition is not as common as lenition. For instance, there are no examples showing the change of a glottal stop to an oral stop.

2.2.1.2 Sound loss

Apheresis. In this sound change, the initial sound in a word is lost. An example of such change can be found in a South-Central Dravidian language, Pengo: *rācu* 'snake' < **trācu*.
Apocope. A sound is lost in the word-final segment in this sound change. An example is: French *lit* > /li/ 'bed'.
Syncope. A sound is lost from the middle of a word. For instance, Old Indo-Aryan *paṭṭa* 'slab, tablet' ∼ Vedic Sanskrit *pattra-* 'wing/feather' (Masica 1993: 157).
Cluster reduction. In this change a complex consonant cluster is reduced to a single consonant. For instance, the initial consonant clusters in English are simplified through the loss of *h*; *hring* > *ring*, *hnecca* > *neck* (Bloomfield 1935: 370). Modern Telugu lost the initial consonant when the initial consonant cluster was of the form *Cr*. Thus *Cr* > *r* : *vrāyu* > *rāyu* 'write' (Krishnamurti and Emeneau 2001: 317).
Haplology. When a sound or group of sounds recur in a word, then one of the occurrence is dropped from the word. For instance, the Latin word *nūtrix* which should have been *nūtri-trix* 'nurse', regular feminine agent-noun from *nūtriō* 'I nourish' where *tri* is dropped in the final form. A similar example is Latin *stipi-pendium* 'wage-payment' > *stipendium* (Bloomfield 1935: 391).

2.2.1.3 Sound addition

Excrescence. This refers to the insertion of a consonant between two consonants. For instance, Spanish developed a [b] in *hombre* 'man' from Latin *hominem* (Crowley and Bowern 2009: 31).
Epenthesis. The insertion of a vowel into a middle of a word. Tamil inserts a

[4] A reconstructed form is indicated by a *.

vowel in complex consonant cluster such as *paranki* < *franco* 'French man, foreigner' (Krishnamurti 2003: 478).
Prothesis. A vowel is inserted at the beginning of a word. Since Tamil phonology does not permit liquids *r, l* to begin a word, it usually inserts a vowel of similar quality of that of the vowel present in the successive syllable when there would otherwise be an initial liquid. Tamil *aracan* < *rājan* 'king' (Krishnamurti 2003: 476).

2.2.1.4 Metathesis

Two sounds swap their position in this change. Latin *miraculum* > Spanish *milagro* 'miracle' where the liquids *r, l* swapped their positions (Trask 2000: 211).

2.2.1.5 Fusion

In this change, two originally different sounds become a new sound where the new sound carries some of the phonetic features from the two original sounds. For instance, compensatory lengthening is a kind of fusion where after the loss of a consonant, the vowel undergoes lengthening to compensate for the loss in space (Crowley and Bowern 2009). Hindi *āg* < Prakrit *aggi* 'fire' is an example of compensatory lengthening.

2.2.1.6 Assimilation

In this type of sound change, a sound becomes more similar to the sound preceding or following it. In some cases, a sound becomes exactly the same as the sound next to it – *complete assimilation*; otherwise, it copies some of the phonetic features from the adjacent sound to develop into a intermediary sound – *partial assimilation*. The Prakrit forms in Indo-Aryan show complete assimilation from their Sanskrit forms: *agni* > *aggi* 'fire', *hasta* > *hatta* 'hand', and *sarpa* > *sappa* 'snake' (B. Lakshmi Bai, p.c.). Palatalization is a type of assimilation where a consonant preceding a front vowel develops a palatal feature, such as [k] > [c]. For example, Telugu shows palatalization from PD: Telugu *cēyi* 'hand'< **key* < **kay* (Krishnamurti 2003: 128).

2.2.1.7 Dissimilation

This type of sound change is opposite to that of assimilation. A classic case of dissimilation is Grassmann's law in Sanskrit and Ancient Greek, which took place independently in the two languages. Grassmann's law states that whenever two sequential syllables had an aspirated stop, the first syllable lost the aspiration. For example, Ancient Greek *thriks* 'hair' (nominative), *trikhos* (genitive) as opposed to **thrikhos* (Trask 2000: 142).

2.2.1.8 Some important sound changes

This subsection deals with some identified sound changes from the Indo-European family and the Dravidian family. These sound changes are quite famous and were originally postulated as *laws*, i.e. *exceptionless* patterns of development. However, there were exceptions to these sound laws which made them recurrent but not exceptionless. The apical displacement is an example of such a sound change in a subset of South Dravidian II languages which is on-going and did not affect many of the lexical items suitable for sound change (Krishnamurti 1978). In apical displacement, non-nasal apical consonants (*ṭ, t̲, l, l̤, z̤, r*), which do not occur word-initially in Proto-Dravidian, moved to the word-initial position.

One of the first discovered sound changes in the IE family is *Grimm's law*. Grimm's law deals with the sound change which occurred in all languages of Germanic branch. The law states that in the first step, the unvoiced plosives became fricatives. In the second step, the voiced aspirated plosives in PIE lost their aspiration to become unaspirated voiced plosives. In the third and final step, the voiced plosives became unvoiced plosives (Collinge 1985: 63). Cognate forms from Sanskrit and Gothic illustrate how Grimm's law applies to Gothic, while the Sanskrit forms retain the original state of affairs:

- C {-Voicing, -Aspiration} ∼ C {+Continuant}: *traya-* ∼ *θreis* 'three'
- C {+Voicing, +Aspiration} ∼ C {+Voicing,-Aspiration}: *madhya-* ∼ *midjis* 'middle'
- C {+Voicing, -Aspiration} ∼ C {-Voicing, -Aspiration}: *daśa-* ∼ *taihun* 'ten'

However, there were exceptions to this law: whenever the voiceless plosive did not occur in the word-initial position or did not have an accent in the previous syllable, the voiceless plosive became voiced. This is known as *Verner's*

law. Some examples of this law are: Sanskrit *pitár* ∼ Old English *faedar* 'father', Sanskrit *(va)vrtimá* ∼ Old English *wurdon* 'to turn'. The importance of Verner's discovery is that it introduces the phenomenon of conditioned sound change.

Another important sound change in the IE linguistics is Grassmann's law. As mentioned above (cf. section 2.2.1.8), Grassmann's law (GL) states that whenever two syllables with aspirated stops (within the same root or when reduplicated) are adjacent to each other, the first syllable's aspirated stop loses the aspiration. According to Collinge (1985: 47), GL is the most debated of all the sound changes in IE. Grassmann's original law has a second proposition regarding the Indic languages where a root with a second aspirated syllable can shift the aspiration to the preceding root (also known as aspiration throwback) when followed by a aspirated syllable. Grassmann's first proposition is mentioned as a law whereas the second proposition is usually omitted from historical linguistics textbooks.

Bartholomae's law (BL) is a sound change which affected Proto-Indo-Iranian roots. This law states that whenever a voiced aspirated consonant is followed by a voiceless consonant, there is an assimilation of the following voiceless consonant and deaspiration in the first consonant. For instance, in Sanskrit, *lab^h+ta* > *labd^{h}a* 'sieze', *dah (< *dag^h) +ta* > *dagd^{h}a* 'burnt', *bud^h+ta* > *budd^{h}a* 'awakened' (Trask 2000: 38). An example of the application of BL and GL is: *budd^{h}a* can be explained as PIE **b^{h}ewd^h* (e-grade) $\overset{GL}{\rightarrow}$ Sanskrit *bud^h* (Ø-grade); *bud^h+ta* $\overset{BL}{\rightarrow}$ *budd^{h}a* 'awakened' (Ringe 2006: 20).

Another well-known sound change in the Indo-European family is umlaut (metaphony). In this change, a vowel transfers some of its phonetic features to its preceding syllable's vowel. This sound change explains singular : plural forms in Modern English such as *foot* : *feet* and *mouse* : *mice*. Trask (2000: 352–353) lists three umlauts in the Germanic branch:

- *i*-umlaut fronts the preceding syllable's vowel when present in a plural suffix in Old English *-iz*.
- *a*-umlaut lowers the vowels [i] > [e], [u] > [o].
- *u*-umlaut rounds the vowels [i] > [y], [e] > [ø], [a] > [æ].

Kannada, a Dravidian language, shows an umlaut where the mid vowels became high vowels in the eighth century: [e] > [i] and [o] > [u], when the next syllable has [i] or [u]; Proto-South Dravidian **keṭu* > Kannada *kiḍu* 'to perish' (Krishnamurti 2003: 106).

2.2.1.9 Computational modeling of sound change

Biologists compare sequential data to infer family trees for species (Gusfield 1997; Durbin et al. 2002). As noted before, linguists primarily work with word lists to establish the similarities and differences between languages to infer the family tree for a set of related languages. Identification of synchronic word forms descended from a proto-language plays an important role in comparative linguistics. This is known as the task of "Automatic cognate identification" in LT literature. In LT, the notion of cognates is useful in building LT systems such as sentence aligners that are used for the automatic alignment of sentences in the comparable corpora of two closely related languages. One such attempt by Simard, Foster and Isabelle (1993) employs similar words[5] as pivots to automatically align sentences from comparable corpora of English and French. Covington (1996), in LT, was the first to develop algorithms for cognate identification in the sense of historical linguistics.[6] Covington (1996) employs phonetic features for measuring the change between cognates.

Levenshtein (1966) computes the distance between two strings as the minimum number of insertions, deletions and substitutions to transform a source string to a target string. The algorithm is extended to handle methathesis by introducing an operation known as "transposition" (Damerau 1964). The Levenshtein distance assigns a distance of 0 to identical symbols and assigns 1 to non-identical symbol pairs. For instance, the distance between /p/ and /b/ is the same as the distance between /f/ and /æ/. A linguistic comparison would suggest that the difference between the first pair is in terms of voicing whereas the difference between the second pair is greater than the first pair. The Levenshtein distance (LD) also ignores the positional information of the pair of symbols. The left and right context of the symbols under comparison are ignored in LD. Researchers have made efforts to overcome the shortcomings of LD in direct as well as indirect ways.

In general, the efforts to make LD (in its plainest form henceforth referred to as "vanilla LD") sensitive to phonetic distances is achieved by introducing an extra dimension to the symbol comparison. This is accomplished in two steps:

1. Represent each symbol as a vector of phonetic features.

[5] Which they refer to as "cognates", even though borrowings and chance similarities are included.

[6] Grimes and Agard (1959) use a phonetic comparison technique for estimating linguistic divergence in Romance languages.

2. Compare the vectors of phonetic features belonging to the dissimilar symbols using Manhattan distance, Hamming distance or Euclidean distance.

A phonetic feature can be represented as a binary value or a value on a continuous (Kondrak 2002a) or ordinal (Grimes and Agard 1959) scales. An ordinal scale implies an hierarchy in the phonetic features – place and manner of articulation. Heeringa (2004) uses a binary feature-valued system to compare Dutch dialects. Rama and Singh (2009) use the phonetic features of the Devanagari script to measure the language distances between ten Indian languages.

The sensitivity of LD can also be improved based on the symbol distances derived from empirical data. In this effort, originally introduced in dialectology (Wieling, Prokić and Nerbonne 2009), the observed frequencies of a symbol-pair is used to assign an importance value. For example, a sound correspondence such as /s/ ∼ /h/ or /k/ ∼ /c/ is observed frequently across the world's languages (Brown, Holman and Wichmann 2013). However, historical linguists prefer natural yet less common-place sound changes to establish subgroups. An example of natural sound change is Grimm's law described in previous subsection. In this law, each sound shift is characterized by the loss of a phonetic feature. An example of an unnatural and explainable chain of sound changes is the Armenian *erku* from PIE **dw-* (cf. section 2.3.1.1). A suitable information-theoretic measure such as Point-wise Mutual Information (PMI) – which discounts the commonality of a sound change – is used to compute the importance for a particular symbol-pair (Jäger 2013; Rama, Kolachina and Kolachina 2013).

List (2012) applies a randomized test to weigh the symbol pairs based on the relative observed frequencies. His method is successful in identifying cases of regular sound correspondences in English ∼ German where German shows changed word forms from the original Proto-Germanic forms due to the High German consonant shift. We are aware of only one effort (Rama, Kolachina and Kolachina 2013) which incorporates both frequency and context into LD for cognate identification. Their system recognizes systematic sound correspondences between Swedish and English such as /sk/ in *sko* 'shoe' ∼ /ʃ/.

An indirect sensitization is to change the transcription of a word from fine-grained IPA to coarser representation such as the sound classes approach of Dolgopolsky 1986. Dolgopolsky (1986) designed a sound class system based on the empirical data from 140 Eurasian languages. Brown et al. (2008) devised a sound-class system consisting of 32 symbols and a few post-modifiers to combine the previous symbols and applied vanilla LD to

various tasks in historical linguistics. One limitation of LD can be exemplified by the example of Grassmann's Law. Grassmann's law is a case of distant dissimilation which cannot be retrieved by LD.

There are other string similarity measures such as Dice, Longest common subsequence ratio (Tiedemann 1999), and Jaccard's index. Dice and Jaccard's index are related measures which can handle long-range assimilation/dissimilation. Dice counts the common number of bigrams between the two words. Hence, bigrams are the units of comparison in Dice. Since bigrams count successive symbols, they can be replaced with more generalized skip-grams which count n-grams of any length and any number of skips. In some experiments whose results are not presented here, skip-grams perform better than bigrams in the task of cognate identification.

The Needleman-Wunsch algorithm (Needleman and Wunsch 1970) is the similarity counterpart of Levenshtein distance. Eger (2013) proposes context and PMI-based extensions to the original Needleman-Wunsch algorithm for the purpose of letter-to-phoneme conversion for English, French, German, and Spanish.

2.2.2 Semantic change

Semantic change characterizes the change in the meaning of a linguistic form. Although textbooks (Campbell 2004; Crowley and Bowern 2009; Hock and Joseph 2009) usually classify semantic change under the change of meaning of a lexical item, Fortson (2003) suggests that semantic change should also be understood to include lexical change and grammaticalization. Trask (2000: 300) characterizes semantic change as one of the most difficult changes to identify. Lexical change includes introduction of new lexical items into language through the processes of borrowing (copying), internal lexical innovation, and shortening of words (Crowley and Bowern 2009: 205–209). Grammaticalization is defined as the assignment of a grammatical function to a previously lexical item. Grammaticalization is usually dealt with under the rubric of syntactic change. Similarly, structural change such as basic word order change, morphological type or ergativity vs. accusativity is also included under syntactic change (Crowley and Bowern 2009; Hock and Joseph 2009).

2.2.2.1 Typology of semantic change

The examples in this section come from Luján 2010 and Fortson 2003 except for the Dravidian example which is from Krishnamurti 2003: 128.

1. *Broadening and narrowing*. A lexical item's meaning can undergo a shift to encompass a much wider range of meaning. Originally, English *dog* meant a particular breed of dog and *hound* meant a generic dog. The word *dog* underwent a semantic change to mean not a particular breed of dog but any dog. Inversely, the original meaning of *hound* changed from 'dog' to 'hunting dog'. The original meaning of *meat* is 'food' in the older forms of English. This word's meaning has now changed to mean only 'meat' and the old meaning still survives in expressions such as *sweetmeat* and *One man's meat is another man's poison*. In another example of narrowing, Tamil *kili* 'bird' $\sim$ Telugu *chili-* 'parrot', Tamil preserves the original meaning and Telugu shifted the meaning to mean a particular bird.

2. *Melioration and pejoration*. In pejoration, a word with non-negative meaning acquires a negative meaning. For instance, Old High German *diorna/thiorna* 'young girl' > Modern High German *dirne* 'prostitute'. Melioration is the opposite of pejoration where a word acquires a more positive meaning than its original meaning. For instance, the original English word *nice* 'simple, ignorant' > 'friendly, approachable'.

3. *Metaphoric extension*. In this change, a lexical item's meaning is extended through the employment of a metaphor such as body parts: *head* 'head of a mountain', *tail* 'tail of a coat'; heavenly objects: *star* 'rock-*star*'; resemblance to objects: *mouse* 'computer mouse'. This change involves similarity.

4. *Metonymic extension*. The original meaning of a word is extended through a relation to the original meaning. The new meaning is somehow related to the older meaning such as Latin *sexta* 'sixth (hour)' > Spanish *siesta* 'nap', Sanskrit *ratha* 'chariot' $\sim$ Latin *rota* 'wheel'. This change involves contiguity.

2.2.2.2 Lexical change

Languages acquire new words through the mechanisms of *borrowing* and *neologisms*. Borrowing is broadly categorized into lexical borrowing (loanwords) and loan translations. Lexical borrowing involves introduction of

a new word from the donor language to the recipient language. Examples of such borrowings are the word *beef* 'cow' from Norman French. Although English had a native word for cow, the meat was referred to as beef and was subsequently internalized into the English language. English borrowed a large number of words through cultural borrowing. Examples of such words are *chocolate, coffee, juice, pepper*, and *rice*. The loanwords are often modified to suit the phonology and morphology of the recipient language. For instance, Dravidian languages tend to deaspirate the aspirate sounds in the loanwords borrowed from Sanskrit: Tamil *mētai* < Sanskrit *mēd^{h}ā* 'wisdom' and Telugu *kata* < Sanskrit *kat^{h}a* 'story'.

Meanings can also be borrowed into a language and are known as *calques*. For instance, Telugu borrowed the concept of *black market* and translated it as *nalla bajāru*. Neologisms is the process of creating new words to represent hitherto unknown concepts – *blurb, chortle*; from person/tribe names – *volt, ohm, vandalize* (from Vandals); place names – Swedish *persika* 'peach' < Persia; from compounding – *braindead*; amalgamation – *altogether, always, however*; from clipping – *gym* < *gymnasium, bike* < *bicycle*, and *nuke* < *nuclear*; from derivation – *print* > *printer, wait* > *waiter*.

2.2.2.3 Grammatical change

Grammatical change is a cover term for morphological change and syntactic change taken together. Morphological change is defined as change in the morphological form or structure of a word, a word form or set of such word forms (Trask 2000: 139–40, 218). A sub-type of morphological change is remorphologization where a morpheme changes its function from one to another. A sound change might effect the morphological boundaries in a word causing the morphemes to be reanalysed as different morphemes from before. An example of such change is English *umlaut* which caused irregular singular : plural forms such as *foot* : *feet*, *mouse* : *mice*. The reanalysis of the morphemes can be extended to words as well as morphological paradigms resulting in a restructuring of the morphological system of the language. The changes of extension and leveling are traditionally treated under analogical change (Crowley and Bowern 2009: 189–194).

Syntactic change is the change of syntactic structure such as the word order (markedness shift in word-order), morphological complexity (from inflection to isolating languages), verb chains (loss of free verb status to pre- or post-verbal modifiers), and grammaticalization. It seems quite difficult to draw a line between where a morphological change ends and a syntactic

change starts.[7] Syntactic change also falls within the investigative area of linguistic typology. Typological universals act as an evaluative tool in comparative linguistics (Hock 2010: 59). Syntactic change spreads through diffusion/borrowing and analogy. Only one syntactic law has been discovered in Indo-European studies called Wackernagel's law, which states that enclitics originally occupied the second position in a sentence (Collinge 1985: 217).

2.2.2.4 Computational modeling of semantic change

The examples given in the previous section are about semantic change from an earlier form of the language to its current form. The Dravidian example of change from Proto-Dravidian **kil-i* 'bird' > Telugu *ciluka* 'parrot' is an example of a semantic shift (narrowing) which occurred in a daughter language (Telugu) from the Proto-Dravidian's original meaning of 'bird'.

The work of Kondrak 2001, 2004, 2009a attempts to quantify the amount of semantic change in four Algonquian languages. Kondrak used Hewson's Algonquian etymological dictionary (Hewson 1993) to compute the phonetic as well as semantic similarity between the cognates of the four languages. Assuming that the languages under study have their own comparative dictionary, Kondrak's method works at three levels:

- *Gloss identity*. Whenever two word forms in the dictionary have identical meanings, the word forms get a semantic similarity score of 1.0.
- *Keyword identity*. In this step, glosses are POS-tagged with an existing POS-tagger and only the nouns (*NN* tagged) are supposed to carry meaning.
- *WordNet similarity*. In this step, the keywords identified through the previous step are compared through the WordNet structure (Fellbaum 1998). The sense distance is computed using a semantic similarity measure such as Wu-Palmer's measure, Lin's similarity, Resnik Similarity, Jiang-Conrath distance, and Leacock-Chodorow similarity (Jurafsky and Martin 2000: chapter 20.6).

The above procedure of computing semantic distance is combined with a phonetic similarity measure called ALINE (Kondrak 2000). The combination of phonetic and semantic similarities is shown to perform better than the individual similarity measures. There are a few other works to compute

[7] Fox (1995: 111) notes that "there is so little in semantic change which bears any relationship to regularity in phonological change".

semantic distance between languages based on bilingual dictionaries (Cooper 2008; Eger and Sejane 2010).

Kondrak assumes that meaning change is restricted to nouns only. In contrast, comparative linguists compare and reconstruct bound morphemes and their functions. Kondrak's algorithms require comparative dictionaries as an input, which require a great amount of human effort. This seems to be remedied to a certain extent in the work of Tahmasebi (2013) and Tahmasebi and Risse (2013), who work with texts and not with comparative dictionaries.

Unlike Kondrak, Tahmasebi works on the diachronic texts of a single language. Tahmasebi's work attempts at identifying the contents and interpreting the context in which the contents occur. This work identifies two important semantic changes, namely *word sense change* and *named entity change*. Automatic identification of toponym change is a named entity related task. An example of named entity change is the reversal of city and town names, in Russia after the fall of Soviet Union, to their early or pre-revolutionary era names such as *Leningrad* > *St. Petersburg* (also *Petrograd* briefly); *Stalingrad* (earlier *Tsaritsyn*) > *Volgograd*.

2.3 How do historical linguists classify languages?

Historical linguists classify languages through comparison of related languages based on diagnostic evidence. The most important tool in the toolkit of historical linguists is the comparative method. The comparative method works through the comparison of vocabulary items and grammatical forms to identify the systematic sound correspondences (cf. sections 2.2.1 and 2.2.2 for a summary of sound change and semantic change) between the languages and then project those sound correspondences to intermediary ancestral languages and further back, to a proto-language. The comparative method reconstructs the phonemes (phonological system), morphemes (morphological system), syntax, and meanings in the intermediary ancestral languages – such as Proto-Germanic. These intermediary languages are then used to reconstruct the single ancestral language such as Proto-Indo-European. The comparative method also identifies the *shared innovations* (sound changes which are shared among a subset of related languages under study) to assign a internal structure (*a branching structure*) to the set of related languages. This task comes under the label of *subgrouping*. Overall, the application of the comparative method results in the identification of relations between languages and an assignment of tree structure to the related languages. However, the comparative method is not without problems. The comparative method works by following the traces left

by the processes of language change. Unlike biology the traces of the earlier language changes might be covered or obliterated by temporally recent changes. Thus the comparative method will not be able to recover the original forms whenever the change did not leave a trace in the language. Also, the comparative method (Harrison 2003) does not work for recovering temporally deep – greater than 8000 years (Nichols 1992) – language change.

2.3.1 Ingredients in language classification

The history of the idea of language relationships, from the sixteenth and seventeenth centuries is summarized by Metcalf (1974: 251) (from Hoenigswald 1990: 119) as follows:

> First, [...] there was "the concept of a no longer spoken parent language which in turn produced the major linguistic groups of Asia and Europe." Then there was [...] "a concept of the development of languages into dialects and of dialects into new independent languages." Third came "certain minimum standards for determining what words are borrowed and what words are ancestral in a language," and, fourth, "an insistence that not a few random items, but a large number of words from the basic vocabulary should form the basis of comparison" [...] fifth, the doctrine that "grammar" is even more important than words; sixth, the idea that for an etymology to be valid the differences in sound – or in "letters" – must recur, under a principle sometimes referred to as "analogia".

The above quote stresses the importance of selection of basic vocabulary items for language comparison and superiority of grammatical evidence over sound correspondences for establishing language relationships.

2.3.1.1 Three kinds of evidence

Meillet (1967: 36) lists three sources of evidence for positing language relationships: sound correspondences obtained from phonology, morphological correspondences, and similarities in basic vocabulary. Basic lexical comparison precedes phonological and morphological evidence during the process of proposal and consolidation of language relationships.

Campbell and Poser (2008: 166) insist on the employment of basic vocabulary for lexical comparison. Curiously, the notion of basic vocabulary was not established on empirical grounds. Basic vocabulary is usually understood to consist of terms for common body parts, close kin,

astronomical objects, numerals from one to ten, and geographical objects. The strong assumption behind the choice of basic vocabulary is that these vocabulary items are very resistant to borrowing, lexical replacement, and diffusion and hence, show the evidence of a descent from a common ancestor. However, basic vocabulary can also be borrowed. For instance, Telugu borrowed lexical items for 'sun', 'moon', and 'star' – *sūrya*, *candra*, and *nakshatra* – from Indo-Aryan languages, and the original Dravidian lexemes – *enḍa*, *nela*, and *cukka* – became less frequent or were relegated to specific contexts. Brahui, a Dravidian language surrounded by Indo-Aryan languages, also borrowed quite a large number of basic vocabulary items.

The second evidence for language relationship comes from sound correspondences. Sound correspondences should be recurrent and not sporadic. The sound correspondences should recur in a specific linguistic environment and not be one-time changes. There should be a regularity when reconstructing the order of sound change which occurred in a daughter language from its ancestral language. For instance, Armenian *erku* 'two' is shown to be descended from PIE **dw-*: **dw-* > **tg-* > **tk-* > **rk-* > *erk-* (Hock and Joseph 2009: 583–584). Usually, cognates are phonetically similar and the sound change which caused the reflex is not a series of sound shifts.

The third evidence for language relationship comes from morphology. A comparison of the copula "to be" across different IE branches is shown in table 2.2. The table shows how the morphological ending for 3rd pers. sg. **-ti* and 1st pers. sg. **-mi* shows similarities across the languages.

Lang.	3rd pers. sg.	3rd pers. pl.	1st pers. sg.
Latin	est	sunt	sum
Sanskrit	ásti	sánti	asmi
Greek	esti	eisi	eimi
Gothic	ist	sind	am
Hittite	ešzi	ašanzi	ešmi
PIE	*es-ti	*s-enti	*es-mi

Table 2.2: A comparison of copula across different IE branches (from Campbell and Poser 2008: 181).

It would be worth noting that the morphological analysis shown in table 2.2 is done manually by reading the texts of these dead languages. In LT, reliable morphological analyzers exist only for a handful of languages and any attempts at an automatic and unsupervised analysis for the rest of the world's languages has a long way to go (Hammarström and Borin 2011).

2.3.1.2 Which evidence is better?

Morphological evidence is the strongest of all the three kinds of evidence to support any proposal for genetic relationships (Poser and Campbell 1992). For instance, Sapir proposed that Yurok and Wiyot, two Californian languages, are related to the Algonquian language family based on grammatical evidence. This claim was considered controversial at the time of the proposal but was later supported through the work of Haas 1958. In the same vein, languages such as Armenian, Hittite, and Venetic were shown to be affiliated to IE based on morphological evidence. Armenian is a special case where the language was recognized as IE and related to Iranian based on lexical comparison. Eventually, grammatical comparison showed that Armenian borrowed heavily from Iranian questioning the earlier conclusion that Armenian is a Iranian language. The grammatical comparison also showed that Armenian is a distinct subgroup within the IE family. When working with all three kinds of evidence the linguist seeks to eliminate borrowings and other spurious similarities when consolidating new genetic proposals. In a computational study involving the ancient languages of the IE family, Nakhleh et al. (2005) perform experiments on differential weighting of phonological, morphological, and lexical characters to infer the IE family tree. They find that weighting improves the match of the inferred tree with the known IE tree. Kolachina, Rama and Bai (2011) apply the maximum parsimony method to hand-picked features in the Dravidian family to evaluate the binary vs. ternary splitting hypotheses at the top-most node.

2.3.2 The comparative method and reconstruction

The comparative method has been described in various works by Hoenigswald (1963, 1973, 1990, 1991), Durie and Ross (1996), and Rankin (2003). The flowchart in figure 2.1 presents an algorithmic representation of the steps involved in the comparative method.

Comparison of basic vocabulary constitutes the first step in the comparative method. In this step, the basic word forms are compared to yield a list of sound correspondence sets. The sound correspondences should be recurring and not an isolated pair such as Greek /t^h/ $\sim$ Latin /*d*/ in *theos* $\sim$ *deus* (Fox 1995: 66) – we know that Greek /t^h/ should correspond to Latin /*f*/ in the word-initial position. These sound correspondences are then used to search for plausible cognates across the languages. Meillet requires that a cognate set should occur in at least three languages to deem the cognate set as plausible. In the next step, a possible proto-phoneme for a sound correspondences set is posited. For

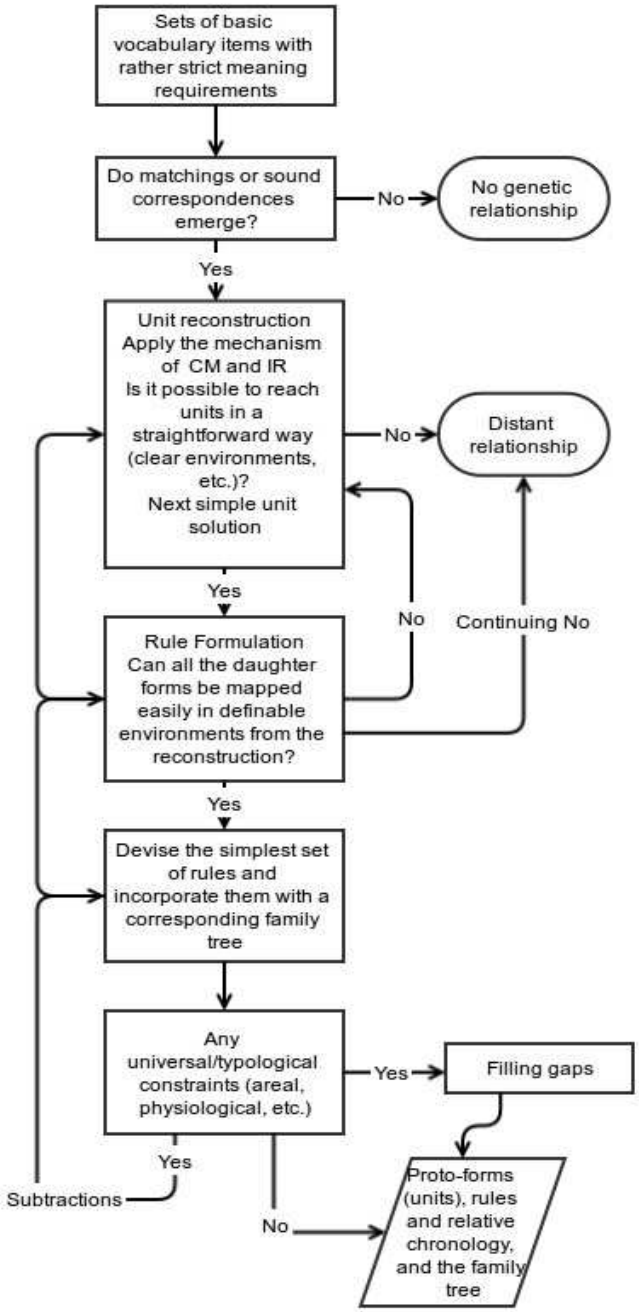


Figure 2.1: Flowchart of the reconstruction procedure (Anttila 1989: 347). CM and IR stand for the comparative method and internal reconstruction.

instance, if a sound correspondence set is of the form *p/p/p*, in the Latin, Greek, and Sanskrit words for 'father', then the proto-phoneme is posited as **p*. In the next step, a phonetic value is assigned to the proto-phoneme. The case of *p/p/p* is a relatively easy one whereas the case of Latin *formus*, Greek t^h*ermos*, and Sanskrit g^h*armas* 'warm' is a recurring sound correspondence of $f/t^h/g^h$. In this case, a consensual phonetic value is assigned to the proto-phoneme. The actual reconstructed proto-phoneme in this case is $*g^{wh}$. This reconstruction comes at a later stage when the proto-phonemes of natural type are established. For instance, even when Armenian *erk-* regularly corresponds to Sanskrit *dw-* in word-initial position, the explanation for such regularity is left for the later stage. Anttila (1989) calls such regular yet non-gradual similarity an evidence for distant relationship. It has to be noted that the assigned phonetic value of a proto-phoneme should not be of any arbitrary value but something that explains the gradual phonetic shift and the change from a proto-phoneme to reflexes should be explainable through the least number of most natural changes.

As noted earlier, regular morphological correspondence provides the strongest evidence for genetic relationship. In fact, Meillet (translated by Poser and Campbell 1992) holds that regular sound correspondences are not absolute proof of relatedness and goes on to stress that irregular grammatical forms are the best evidence for establishing a *common language*, i.e., a proto-language. According to Anttila (1989), what passes as morphological reconstruction is mostly phonological in nature (*morphophonemic analysis*). Morphophonemic reconstruction makes up the reconstruction of grammatical forms and their grammatical function.

The reconstruction of the lexicon or the meaning of the reconstructed proto-forms is not parallel to that of phonological reconstruction. According to Fox (1995: 111–118), the lexicon reconstruction procedure does not have the parallel step of positing a proto-meaning. The next step after the comparison of daughter languages' meanings is the reconstruction of the proto-meanings. An example of such reconstruction are the assignment of meaning to the IE proto-form **pont*. Greek has two meanings 'sea' and 'path'; Latin and Armenian have the meanings 'ford' and 'bridge'; Sanskrit and Old Church Slavonic have the meanings of 'road' or 'path'. Vedic has the meaning of 'passage' through air as well. A reconciliation of these different meanings would indicate that the original form had the meaning of 'passage' which was extended to 'sea' in Greek, a narrowing to travel over water or land in Latin and Armenian. So, the original meaning of **pont* is reconstructed as a general word for travel. To conclude, the lexicon reconstruction is done on a per-word basis and is not as straightforward as phonological reconstruction.

Typological universals serve as a sanity check of the reconstructed languages' linguistic systems. For instance, positing an unbalanced vowel or

consonant system would be untenable under known typological universals. Hock (2010: 60) summarizes the 'glottalic' theory in Indo-European languages as an example of typological check of the reconstructed consonant system. The PIE consonant inventory has a voiceless, voiced, and voiced aspirate consonants. This system was asserted as typologically impossible since any language with voiced aspirates should also have voiceless aspirates. A glottalized consonant series in addition to the voiceless aspirates was proposed as the alternate reconstruction that satisfies the conditions imposed by typology. Working from PIE to the daughter languages, the expanded consonant system would make Grimm's law unnecessary and suggests that the Germanic and Armenian consonant systems preserve the original PIE state and all the other IE languages have undergone massive shifts from PIE. The glottalic system has lost its importance after the discovery of Indonesian languages which have voiced aspirates without their voiceless counterparts. Moreover the glottalic system is against the general principle of Occam's Razor (Hock and Joseph 2009: 443–445).

The regular sound correspondences established through the comparative method also help in recognizing borrowings. For instance, English has two forms with meanings related to 'brother' *brotherly* and *fraternal*. The regular sound correspondence of PIE $*b^h > b$ suggests that *fraternal* is not a native word and it was in fact borrowed from Latin.

In this step, the enumeration of shared innovations and shared retentions form the next stage for positing a family tree. Shared innovations are regular and natural sound changes shared by a subset of languages. The shared innovations in a subset of languages suggest that these languages have descended from an intermediary common ancestor which has undergone this particular linguistic change and all the daughter languages of the intermediary ancestor show this change. Grimm's law is such a sound change which groups all the Germanic languages under a single node. Meillet (1967: 36) employs a different term *shared aberrancies* (also called *shared idiosyncrasies* by Hock and Joseph 2009: 437) such as the recurrent suppletive form correspondence between English and German for a strong evidence of the genetic relationship.

Despite the copious research in IE linguistics, the tree structure for IE at higher levels is not very well resolved (cf. figure 2.2). A basic assumption of the comparative method is that the proto-language is uniform and without dialectal variation. However, there are reflexes among daughter languages whose correspondence cannot be accounted for from known evidence. In such a case, a practitioner of the comparative method has to admit it as dialectal variation. An example of the admittance of dialectal variation in proto-language is the correspondence of voiceless aspirates in Indo-Iranian to

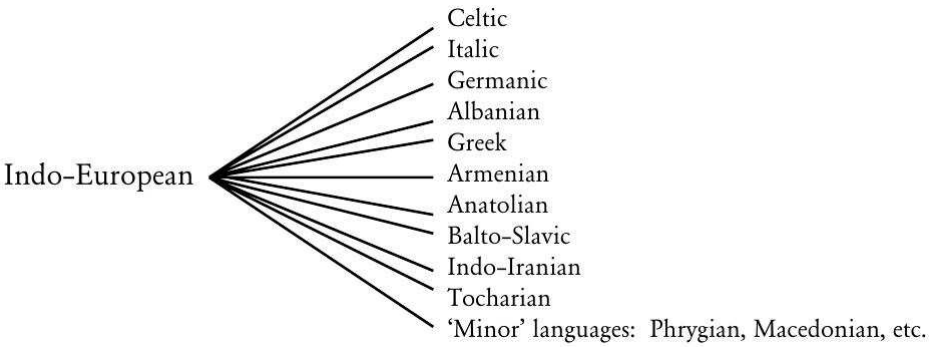


Figure 2.2: Higher-order tree of IE family from Garrett (1999).

other IE branches: Sanskrit *rat^{h}a-* ∼ Latin *rota* 'chariot, wheel'. Finally, the comparative method assumes that sound change operates without exceptions or it affects all the suitable lexical items. However, Krishnamurti (1978) demonstrated a sound change known as apical displacement (cf. section 2.2.1.8) which is still in progress (*lexical diffusion*; Chen and Wang 1975) in few languages of the South Dravidian II family but has proceeded to completion in Gondi. Based on a single innovation which is still in progress, Krishnamurti, Moses and Danforth (1983) infer the family tree for the South-Central Dravidian family using the unaffected cognates as a criterion for subgrouping. In another study, based on the same dataset of South-Central Dravidian languages, Rama, Kolachina and Bai (2009) apply different phylogenetic techniques listed in section 2.4.2 and find that the different phylogenetic methods largely agree with the classification given by the comparative method.

2.3.2.1 Tree model

A tree model only represents the genetic affiliations inside a language family and does not represent the dialectal borrowings and borrowings from neighboring related languages. Also, a parallel (independent) development such as Grassmann's law in Greek and Sanskrit cannot be shown in the tree model. Moreover, the tree resulting from the application of the comparative method is not metrical[8] and does not explicitly show information about the date of splits (Hoenigswald 1987). The date of splits can be worked out

[8] A metrical tree shows branch lengths.

through epigraphic evidence, relative chronology of the sound changes, and archaeological evidence. As Bloomfield (1935: 311) points out:

> The earlier students of Indo-European did not realize that the family-tree diagram was merely a statement of their method; they accepted the uniform parent languages and their sudden and clear-cut splitting, as historical realities.

The above statement suggests that the tree is only a model or device to represent the inherited linguistic characteristics from a common ancestor. Moreover, the comparative method attempts to establish a successive split model of a language family. Also, a family tree obtained through the application of the comparative method need not show binary splits at all the nodes. This can either be due to the lack of information for resolving the unresolved nodes or might represent the actual language diversification history of the language family. For example, the Dravidian family tree shows a ternary split at the root (Krishnamurti 2003: 493).

A mathematical treatment of the enumeration of possible rooted binary vs. non-binary trees is given by Felsenstein (2004: 19–36). The number of possible rooted, non-binary, and unlabeled trees for a given family size is presented in table 2.3.

Family size	Tree shapes
2	1
5	12
10	2312
20	256738751
40	9.573×10^{18}
80	3.871×10^{40}
100	2.970×10^{51}

Table 2.3: Number of non-binary tree topologies.

2.3.2.2 Wave model

The observation that there were similarities across the different branches of the IE family led to the wave model, proposed by Schmidt (1872). The wave model for the IE language family is given in figure 2.3. For instance, the Balto-Slavic, Indo-Iranian, and Armenian subfamilies share the innovation

from original velars to palatals. In this model, an innovation starts out in a speech community and diffuses out to neighboring speech communities. An example of an isogloss map for South Dravidian languages is given in figure 2.4. The wave model is not an alternative to the tree model but captures some regularities not shown by the tree model. The wave model captures the overlapping innovations across the subfamilies and also shows the non-homogeneity of the proto-language. Representing the proto-language at one end and dialects of a daughter language at the other end on a graded scale, the tree model can be reconciled with the wave model. The tree-envelope representation of Southworth 1964 is one such example which attempts to show the subgrouping as well as the shared innovations between the subgroups. The study of lexical diffusion of $s > h > \emptyset$ in Gondi dialects by Krishnamurti (1998) is an example where the original Proto-Dravidian $*c > *s$ in the word-initial, pre-vocalic position completed the sound change in South Dravidian languages. This sound change is succeeded by $*s > *h > \emptyset$ and is completed in South Dravidian I and Telugu. The same sound change is still ongoing in some Gondi dialects and the completion of the sound change marks the dialectal boundary in Gondi.

2.3.2.3 Mesh principle

The mesh principle is developed by Swadesh (1959) for identifying the suspected relations between far-related languages. Swadesh begins by observing that the non-obvious relationship between Tlingit and Athapaskan becomes obvious by including Eyak into the comparative study. In parallel to the situation of a dialectal continuum, there is also a lingual chain where the links in the chain are defined through systematic grammatical and sound correspondences. Swadesh (1959: 9) notes that:

> However, once we have established extensive networks of related languages connected with each other in a definite order of relative affinities, expressible, for example, in a two-dimensional diagram, it is possible to test each new language, as yet unplaced, at scattered points in the constellation to find where it comes the nearest to fitting.

This can be easily related to the Multi-dimensional Scaling technique (MDS; Kruskal 1964) which projects a multi-dimensional matrix to a two-dimensional representation. Consider the task of placing the position of a recalcitrant language in relation to other established subgroups, say Armenian. The first step in this model will create a MDS diagram of IE

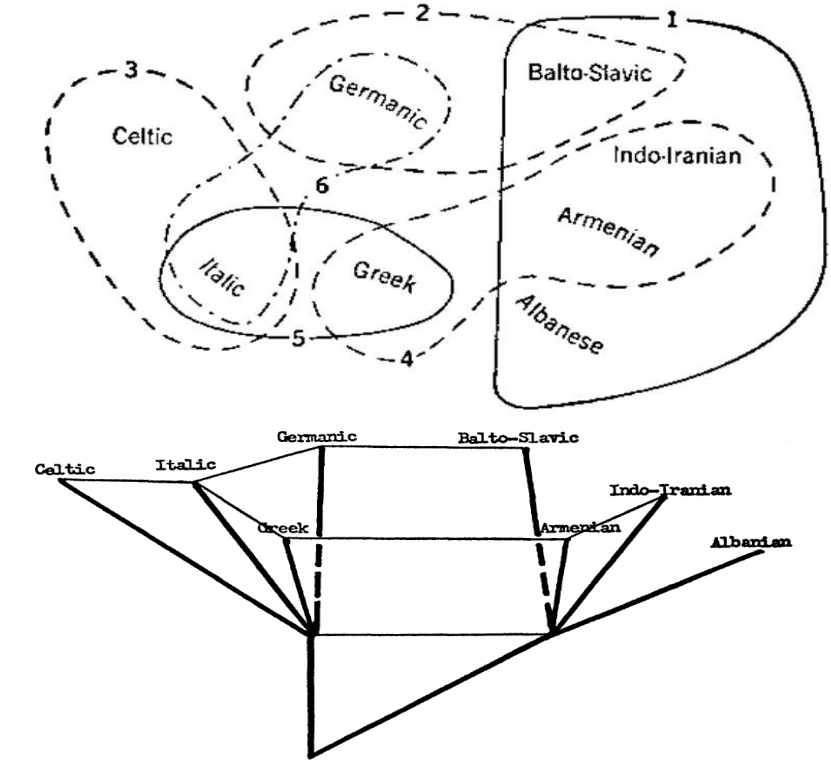
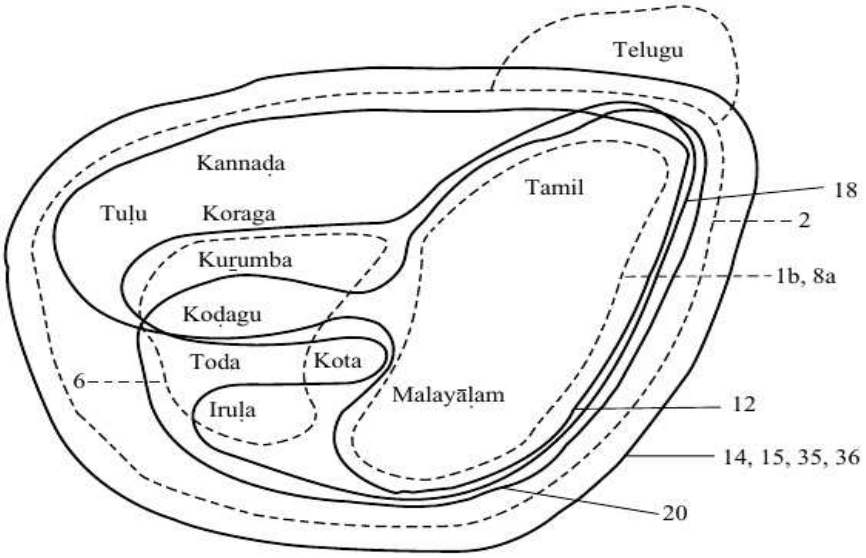


Figure 2.3: Indo-European isoglosses (Bloomfield 1935: 316) and the corresponding tree-envelope representation from Southworth (1964). The numbers in isogloss figure correspond to the following features. **1.** Sibilants for velars in certain forms. **2.** Case-endings with [m] for [b^h]. **3.** Passive-voice endings with [r]. **4.** Prefix [e-] in past tenses. **5.** Feminine nouns with masculine suffixes. **6.** Perfect tense used as general past tense.

languages without Armenian and then repeat the step with Armenian to see the shift in the positions of other languages due to the introduction of Armenian. A much simpler case would be to remove a pivotal language such as Sanskrit – that provided evidence for stress patterns in PIE (cf. Verner's law) – to produce a MDS representation and then repeat the step to see the shift of the languages in the fuller picture.

Given the recent application of biological network software to linguistic data, Nichols and Warnow (2008) divide the mesh-like representations into two categories: implicit and explicit networks. Implicit networks do not show the explicit interaction (such as borrowing and diffusion) between two independent languages such as French and English but show a mass of



Phonological isoglosses

Morphological isoglosses

F1b. PSD *e *o > *i *u/__+a
F2. PD *c > Ø- (through *s- > *h- not attested directly)
F6. Centralized vowels in root syllables
F8a. PD *k > c-/#__V [–Back], C [–Retroflex]
F12. Addition of *-$kaḷ$ (n-hpl suff) optionally to 1pl and 2pl
F14. Creation of *aw-$aḷ$ etc. 3f sg
F15. Loss of *$ṯ$ in 3m sg *aw-$anṯ$, *iw-$anṯ$ 'he'
F18. Loss of -Vn as accusative marker
F20. Tense–voice marking by final
NP ~ NPP
F35. Use of $tān$ 'self' as an emphatic particle; alternatively -$ē$
F36. copular verb *ir- 'be' replacing *man-

Figure 2.4: Shared innovations in South Dravidian I represented as isoglosses (Krishnamurti 2003: 498).

inherited linguistic material at the center of the network. The farther one gets away from the center and towards the branches of the network, the greater linguistic divergence one observes in the daughter languages. An example of such a network drawn from the cognate data of the *Dravidian Etymological Dictionary* (Burrow and Emeneau 1984) is given in figure 2.5. Explicit networks show the contact scenario between the different branches in a family tree and are inferred from the three kinds of evidence (Nakleh, Ringe and Warnow 2005).

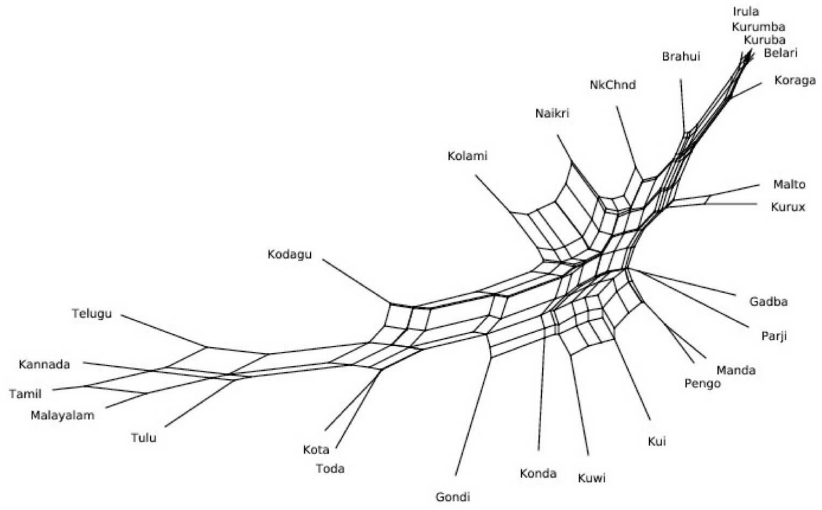


Figure 2.5: A network diagram of 28 Dravidian languages based on grammatical and phonological features (Rama and Kolachina 2013).

2.3.2.4 The comparative method as an iterative technique

The comparative method as explained in the previous section is iterative in nature. The flowchart presented in figure 2.1 captures the iterative aspect of the comparative method. In the initial stages, the method accumulates evidence from basic vocabulary comparison and either reinforces or weeds out putative daughter languages from comparison. Similar to sound change that is characterized to affect the suitable parts of vocabulary so does the comparative method adds more evidence to it as it scans through more linguistic material. The initial set of languages is always based on diagnostic evidence and not grounded in solid evidence. As Nichols (1996) notes, some branches of Indo-European such as Slavic were always known to be related due to the medieval records which were part of the Germanic philological tradition. As the structure of the language family becomes concrete, the remaining proto-language systems are established with evidence from the neighboring daughter languages as well other intermediary ancestors (*inverted reconstruction*; Anttila 1989: 345–346).

The *modus operandi* of the comparative method has parallels in LT. Many LT systems which work in the semi-supervised fashion begin with a seed list of annotated linguistic examples. The seed list is supposedly small and the original LT system is supposed to achieve high accuracy. In the next step,

more unannotated linguistic examples are supplied to the LT system for the classification task and a human annotator judges the performance of the LT system on each unannotated example as correct or incorrect at the end of a step. The correct examples are added back to the original seed list to train the next version of LT system. This process is repeated until there is no increase in the accuracy of the LT system.

Hauer and Kondrak (2011) employs this paradigm to boost a cognate identification system's accuracy by self-learning the language relatedness parameter. SMT systems are another LT parallel to the comparative method. Given a large parallel corpus of two languages with no other linguistic annotation, SMT systems learn the phrase to phrase translations between the language. In the first iteration, any source language phrase can be mapped to target language phrase with equal chance. As the learning proceeds, the probabilities (evidence) for the source-target maps change and reach a local optimum where the evidence does not change over iterations. In a similar fashion, as evidence for language relationship accumulates, the comparative method's earlier predictions are subjected to change.

Bouchard-Côté et al. (2013) reconstruct Proto-Austronesian lexemes from the 200-word Swadesh list of 659 Austronesian languages. They assumed the tree topology of Austronesian language family as given and then proceeded to reconstruct the proto-word forms of the 200 meanings. It has to be noted that their method does not come close to the comparative method as the tree structure is given by linguists and not inferred from the data. These authors reduce the reconstruction step to a search procedure over a tree topology. Hence, there is an inherent circularity in their method.

2.4 Complementary techniques in language classification

The positing of genetic proximity based on cognate counts and the counts of shared phonological and grammatical innovations preceded lexicostatistics. This is noted by Swadesh (1959) where Kroeber in 1907 used the established innovations to draw a two-dimensional proximity maps for Californian languages. Campbell (2004) also makes the point that only a *shared innovation* can be used to classify languages. This brings us to an important question, if there can be any method other than the comparative method to establish subgroups or classify languages.

2.4.1 Lexicostatistics

The lexicostatistical technique as introduced by Swadesh (1950) works on standardized multi-lingual word lists. Contrary to the popular conception that the similarities between two word lists are based on *look-alikes*, two words are judged to be similar if and only if they are cognates. The meanings in these lists are supposed to be resistant to borrowing and internal lexical replacement. The important question is how did Swadesh arrive at such a list? The multiple families studied in CHL show that the list is actually robust and the classifications inferred from the standardized word lists come close to the classifications proposed through the comparative method (Greenhill and Gray 2009; Wichmann et al. 2010a).

The issue of origin is investigated by Tadmor, Haspelmath and Taylor (2010). The authors quote from Swadesh (1971: 19) about the creation and refinement process from 215-word list to 100-word list.

> In counting and statistics, it is convenient to operate with representative samples, that is, a portion of the entire mass of facts so selected as to reflect the essential facts. For our lexical measure of linguistic divergence we need some kind of selected word list, a list of words for which equivalents are found in each language or language variant ...

Apart from using the word lists for glottochronological studies, Swadesh intended to make the 100-word list a *diagnostic vocabulary* for investigating known as well as suspected language relationships.

2.4.2 Beyond lexicostatistics

A large amount of research has been conducted based on the 100 or 200 item word lists. The availability of off-the-shelf biological software spurred researchers to apply the biological methods to the Swadesh word lists to yield family trees based on distance-based methods as well as character-based methods. An excerpt of such input data is given in tables 2.4 and 2.5.

The character encoding technique present in the table 2.5 does not distinguish between shared retentions and shared innovations. The technique does not reconstruct the observed character's ancestral states at the internal nodes of a tree. In this context, a character is synonymous with trait in biology. A character shows how a single trait changes its value as it evolves along the branches of a tree. The linguistics counterpart of a character is a reconstructed linguistic feature such as proto-lexeme which can be lost it evolves along the branches of a tree.

Items	Danish	Swedish	Dutch	English
'person'	menneske/1	människa/1	mens/1	person/2
'skin'	skind/1	skinn/1, hud/2	huid/2	skin/1

Table 2.4: Two lexical characters for four Germanic languages (Wichmann 2010a: 77–78). Each cell corresponds to a word form in a language and its cognacy state. word forms with the same state are cognates.

Items	Danish	Swedish	Dutch	English
'person-1'	1	1	1	0
'person-2'	0	0	0	1
'skin-1'	1	1	0	1
'skin-2'	0	1	1	0

Table 2.5: The binary encoding of the lexical characters given in table 2.4 (Wichmann 2010a: 79).

For example, PIE *ǵʰésr̥ 'hand' is retained in the Anatolian, Greek, Albanian, Tocharian, and Indo-Iranian branches whereas the Balto-Slavic, Germanic, Celtic, and Italic branches show individual innovations (cf. table 2.6). As noted above, the binary encoding does not distinguish retentions from innovations. To remedy this, I propose a new encoding in those cases where the PIE form has been established with general consensus. In these cases each innovation would be treated as a character and those languages that exhibit the reflex of PIE form are given an **R**etention state. For instance,

Languages	Hand	R/I/A encoding				
Hittite	1	R	A	A	A	A
Vedic	1	R	A	A	A	A
Old Church Slavonic	2	A	I	A	A	A
Old High German	3	A	A	I	A	A
Old Irish	4	A	A	A	I	A
Latin	5	A	A	A	A	I

Table 2.6: Multi-state character encoding for 'hand'. Hittite and Vedic show the reflex of PIE and exhibit state '1'. Each innovative state is shown as 'I' in a separate column. The binary encoding is obtained by replacing all R's, I's by '1' and 'A's by '0'.

in the example of 'hand', the Balto-Slavic branch would show a **I** state for its innovation whereas the retention languages show a **R** state and the Germanic, Celtic, and Italic branches would show a **A**bsence state.

In some cases, the PIE language can show two forms for the same meaning. For instance, PIE 'bone' shows two forms *$h_3$ésth_1- (a **R1** in Indo-Iranian) and *kosti- (a **R2** in Slavic) whereas, North West Germanic languages show a innovation. Hence, the Indo-Iranian languages show a **R** state for the Germanic innovation character and Celtic languages show a **A**. This example is the case of variation in the PIE form that resulted in two cognate classes in the shared retentions. PIE daughter languages show two cognate classes for the meaning 'egg' where both the PIE forms are related through ablaut variation. Some meanings such as 'burn' show three different forms at the PIE stage that are reconstructed from multiple daughter languages. The meaning 'burn' would be treated the same as 'bone' but with three retention states – **R1**, **R2**, and **R3** – that can be collapsed into a single state **R**.

In some other cases, the PIE form's cognate class is not available and none of the languages would show a **R** state. For example, PIE 'neck' is reconstructed as *mon- but its cognate class is unknown. Sometimes, the PIE form is simply not available. For example, the PIE root for 'bark' (noun) is not available and as such none of the daughter languages show an **R** state.[9]

The design of the states R, I, and A assumes the following paths of character evolution:

- The root node of the tree has a retention that was lost and replaced by an innovation in one of its descendant branches.
- Or there was an innovation that coexists with the original retention. This might lead to the observed multiple states for a single character. This condition is known as polymorphism and can be problematic for a phylogenetic inference algorithm (Ringe, Warnow and Taylor 2002).
- Or, the state at the root node is unknown; in such a case, the algorithm can probabilistically reconstruct the root node's state.

2.4.2.1 Character-based methods

The character based methods are designed to find the best tree according to a model of character evolution that optimizes a criterion. The model of character evolution is known as the substitution model (explained below) and

[9] All the examples are from the IELex database (cf. section 3.1.3). The 'hand' example is from Ringe, Warnow and Taylor 2002.

the optimization criterion is one of the following: maximizing the likelihood of the data, minimizing the sum of branch lengths, or the least number of state changes (for example, the transitions between states R, I, A). The optimization step is usually performed after fixing the tree topology.

2.4.2.1.1 Substitution models

A substitution model consists of a $n \times n$ matrix (Q) that provides the instantaneous rate of change between the states in an infinitesimal time Δt ($\Delta t \to 0$). If the space of states is R, I, and A, then the Q matrix is of 3×3 size and its rows sum to zero. The Q matrix is used to compute the probability of being in a state j given the original state i after time t. In terms of tree, this computation gives the branch length between two nodes provided we known the states at each end of the branch. Each element in the transition probability matrix is denoted by $P_{ij}(t)$ and is computed as $e^{q_{ij}v}$ where v is the branch length between two nodes in a given tree. The derivation of the branch length formula and the properties of Q matrix can be found in Yang 2014: chapter 1. In reality, the branch lengths and the Q matrix are estimated in an iterative fashion. There are many ways to estimate the branch length and the Q matrix but the most popular way is to model it as a constrained optimization problem. This step requires the concept of likelihood of a tree which I will explain in the next paragraph.

2.4.2.1.2 Tree likelihood

The tree likelihood is defined as the likelihood of observing a character sequence under a fixed tree topology. A character sequence is a row from table 2.5. The computation of the likelihood is achieved through the pruning algorithm of Felsenstein 1981.[10] The pruning algorithm computes the maximum likelihood over all possible states at the internal nodes of the tree.[11]

The constrained optimization problem would be to find the positive branch lengths that maximize the likelihood for a Q matrix that is supplied beforehand. This step can be achieved through Newton-Raphson's method. The next step would be to estimate the Q matrix by fixing the tree's branch lengths and topology. This step can be modeled as a constrained optimization problem where the Q matrix is written as the product of two symmetric

[10]The pruning algorithm is a special case of dynamic programming algorithm on a tree.

[11]This is quite similar to the task of likelihood estimation of an observed word sequence in POS-tagging in LT.

matrices S and Π. The Π matrix is a diagonal matrix whose non-diagonal elements are zero and the diagonal elements are positive and sum to 1. The non-diagonal elements of S matrix are positive. Both the steps are repeated until convergence (Yang 2014: 139). The order of optimization steps can be reversed by estimating the Q matrix followed by the branch lengths.

Until now, the tree topology is assumed to be fixed. In reality, one has to exhaustively explore the complete tree space to find the tree that maximized the likelihood. Due to the large number of trees, even for a dataset of twenty languages, the tree space is explored heuristically through the algorithms given below.

2.4.2.1.3 Heuristic tree search

There are three main tree search algorithms:

1. Nearest neighbor interchange: This algorithm works on internal branches of a tree by randomly selecting a internal branch and swapping a randomly selected subtree on one side to another randomly selected subtree on the other side of the branch.

2. Subtree pruning and regrafting: This algorithm works by pruning a subtree and then attaching it to a randomly selected location on the pruned tree.

3. Tree bisection and reconnection: In this algorithm, a internal branch is randomly selected and removed to break the tree into two subtrees. In the next step, two randomly selected branches from each subtree are joined by a new branch.

There are many other ways to perform tree rearrangement but the above mentioned algorithms are the most popular of all (Lakner et al. 2008).

2.4.2.1.4 Phylogenetic tree inference

There are three phylogenetic tree inference techniques which have been developed over the last fifty years. They are maximum parsimony, maximum likelihood, and Bayesian inference.

- Maximum Parsimony is an evolutionary principle which demands that the best tree for the data is the one which explains the change of character states from ancestral to leaf nodes in the least possible

number of changes. This method employs the above tree heuristics to explore the tree space and find the tree that shows the least number of state changes.

- Maximum likelihood approach was introduced in the previous section and this approach explores both the tree topologies and branch lengths to find the best tree that maximizes the likelihood of the observed data. The likelihood approach is computationally expensive since it requires the repeated estimation of branch lengths and substitution parameters on all the trees proposed through heuristic search. However, clever techniques have been developed to reduce the computational cost of likelihood by caching the subtree likelihoods (Yang 2014: 143).

- The Bayesian approaches have become popular due to the availability of MrBayes software (Ronquist and Huelsenbeck 2003) which has been employed in the Indo-European phylogenetic analysis (Gray and Atkinson 2003; Atkinson et al. 2005; Atkinson and Gray 2006). The Bayesian approach employs a Monte-Carlo Markov Chain (MCMC;Yang 2014: 214) to sample a tree topology, branch lengths, and substitution matrix (known as a chain's state). The algorithm samples states by running a chain of long length (usually in the order of some millions of states). The initial states of the chain are discarded as part of burn-in. The algorithm also drops states within a pre-specified interval to prevent correlation between adjacent states in the chain. At the end of the run, if the chain shows signs of convergence, then the sampled trees and the substitution matrix's parameters are summarized. Otherwise, the chain has to be run for longer periods of time to ensure proper mixing and convergence.

2.4.2.2 Distance-based methods

In contrast to the character-based methods that model character evolution across a tree, distance-based methods employ the distance between a pair of character sequences to fill the pair-wise distance matrix. The number of unmatched character states is given as the observed distance between two languages. The distance-based methods also work with a fixed tree topology but attempt to find the individual branch lengths by minimizing the sum of squared differences between the observed distances and the estimated distances. This estimation procedure is known as ordinary least squares estimation. If each pair-wise distance is assigned a weight then the optimization procedure is known weighted least squares estimation. A much

general procedure called generalized least squares takes the non-independence of the distances into account when estimating the branch lengths on a fixed topology. The tree inference methods such as Neighbor-Joining (NJ) are employed to infer the tree (Felsenstein 2004: chapter 11) in the case of large datasets (of few hundreds of species).

Multilateral comparison is another alternative language classification technique developed by Greenberg (1993). This method consists of visual inspection of large word tables similar to the one in table 2.4. A large number of languages are compared in a single go and similarity between languages are used to propose a subgrouping for the languages. Greenberg's aim was to propose a single super-family for a large number of Eurasian families. His methods have been criticized (Ringe 1992) due to the lack of support of statistical significance.

2.5 A language classification system

The computational modeling of the entirety of the comparative method would require a language classification system which models each step of the comparative method. Steiner, Stadler and Cysouw (2011) propose such a system (cf. figure 2.6) and applies it to the classification of a Caucasian group of languages and some South American languages that figure in the Intercontinental Dictionary Series (Borin, Comrie and Saxena 2013). One can easily see that pairwise alignments are used to build multiple alignments following Meillet's rule of thumb for including at least three languages into comparison. However, multiple alignment of words is not a straightforward task since it is a NP-complete problem. The NP-completeness is circumvented through the use of pair-wise alignments in an iterative or progressive fashion (Durbin et al. 2002: 134–159). The next section summarizes the different tree evaluation techniques and the computation of deviation from tree-likeness (reticulation) in CHL.

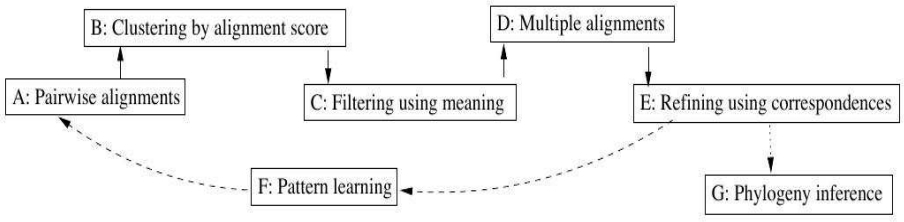


Figure 2.6: A pipeline for a language classification system.

2.6 Tree evaluation

A tree comparison measure quantifies the difference between the family tree inferred from automatic language classification systems and the family tree inferred from the comparative method. The comparative method assumes that languages diverge in a step-by-step fashion yielding a tree. However, it is widely known that language evolution is not always tree-like. For instance, English has borrowed French vocabulary but is still a Germanic language due to its descent from Proto-Germanic. As noted previously, a network model is a graphical device showing the amount of deviation from tree-likeness. But it does not provide a number for the amount of deviation. The δ measure fills in this gap and provides a score for deviation from tree-likeness. The four different tree comparison techniques and δ are described in the following section.

2.6.1 Tree comparison measures

Robinson-Foulds (RF) Distance. The RF distance is defined as the number of dissimilar bipartitions between an inferred tree and gold-standard tree. A bipartition is a pair of language sets resulting from the removal of a internal edge in a phylogenetic tree. For a phylogenetic tree with N languages, there are at most $N-3$ bipartitions. Thus, the RF distance measures the dissimilarity in the topology between the inferred tree and the corresponding family tree. It should be noted that the RF distance does not take branch lengths into account. Any tree inference algorithm yields a phylogenetic tree with branch lengths. RF distance throws away the branch length information when comparing the inferred tree with the family tree. Steel and Penny (1993) introduced three other measures as alternatives to RF distance. Each of these measures are described in detail below.

Branch Score Difference (BSD). BSD is related to RF and takes into account branch lengths. Instead of computing the number of dissimilar partitions between the inferred tree and family tree, BSD computes the sum of the absolute difference in each of the internal branch lengths in the two trees. If an internal branch is absent in one tree and present in the other tree then the branch length for the absent branch is treated as zero.

Path Length Distance (PD). This measure is based on the idea that the distance between two languages can be expressed as the number of edges (branches) in the shortest path (in the tree) connecting the two languages. Each cell of a path length matrix (PDM) consists of the path length between a pair of languages in a phylogenetic tree. PD is computed as the square root of

the average of the square of the difference between each cell of the PDM of the inferred tree and the corresponding cell in the PDM of the linguistic tree.

Weighted Path Length Distance (WPD). WPD is computed in a similar fashion to that of PD except that the path length for a pair of languages, is computed as the sum of the branch lengths of the edges in the path connecting the pair of languages. The WPD matrix (WPDM) is computed similarly to the PD matrix and the WPD is computed as the square root of the average of the square of the difference between each cell of WPDM of the inferred tree versus the family tree.

2.6.2 Beyond trees

Delta (δ). Given a distance matrix d for a language family, δ, the measure of reticulation, is computed as follows:

1. There are $\binom{N}{4}$ quartets for a language family of size N. A quartet, q, is defined as a set of four languages, $\{i, j, k, l\}$. Enumerate all the quartets for a language family.

2. The distance sub-matrix for a quartet can be represented by a tree. If the distances represented in a quartet tree are exactly the same as the distances given in the sub-matrix, then the tree is called *additive*. An example of additive trees is given in figure 2.7.

3. The relation between all the pair-wise distances, in a quartet, can be expressed as follows:

$$d_{ij} + d_{kl} \geq d_{ik} + d_{jl} \geq d_{il} + d_{jk} \tag{1}$$

4. The so-called four point condition is based on (1) and can be expressed as follows:

$$d_{ij} + d_{kl} = d_{ik} + d_{jl} \geq d_{il} + d_{jk} \tag{2}$$

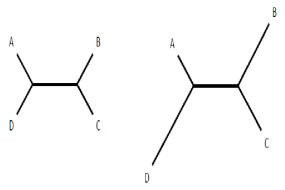


Figure 2.7: Additive trees for a quartet

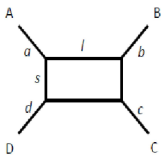


Figure 2.8: Reticulate quartet

Computation: An example of a reticulate quartet is shown in figure 2.8. It carries labels similar to those given in Holland et al. (2002). The labels represent the lengths of each of the 8 edges in the reticulate quartet.

1. The amount of deviation from treelikeness – reticulation – of a quartet can be measured as a deviation from (1).

2. The reticulation measure δ for a quartet is computed as $\delta = \frac{s}{l}$ where, $s = d_{ij} + d_{kl} - d_{ik} - d_{jl}$ and $l = d_{ij} + d_{kl} - d_{il} - d_{jk}$.

3. δ ranges from 0 (when the quartet is additive) to 1 (when the box is a square). The δ for a family is computed through the average of the δ across all the quartets.

4. Wichmann et al. (2011a) suggest the idea of computing the δ for each language in a family but do not pursue this line of investigation further, instead computing δ for few chosen languages only. δ for a language is computed as the average of δs of all the quartets in which a language participates.

Gray, Bryant and Greenhill (2010) compare a related measure of reticulation, *Q-residual* with δ. The reported results are not accurate since the software *SplitsTree* (Huson and Bryant 2006) was discovered to have a bug (Wichmann et al. 2011a).

2.7 Dating and long-distance relationship

Any standard textbook in historical linguistics (Trask 1996; Campbell 2004; Hock and Joseph 2009; Crowley and Bowern 2009) has a chapter on language classification (or relationship) followed by a chapter on *macro-families, proto-world,* and *long-distance relationships*. Only Trask 1996 and Crowley and Bowern 2009 follow the macro-families chapter with a description of statistical techniques employed for assessing the significance of long-distance relationships.

The chapter(s) on language classification consist of the comparative method and its demonstration to a medium-sized language family, such as the Mayan or the Dravidian family. For instance, Campbell (2004) has a chapter on the comparative method and illustrates the use of *shared innovation* in the subgrouping of the Mayan language family. Likewise, Trask (1996) demonstrates the reconstruction of part of Proto-Western Romance vocabulary through the application of the comparative method to synchronic Romance language vocabulary lists. The reconstruction of proto-world is

characterized as a *maverick* approach by Trask. Any quantitative technique which attempts at dating the divergence time of a language into its daughter languages is bundled together with glottochronology.

For instance, Campbell (2004) uses the terms *glottochronology* and *lexicostatistics* interchangeably. Although both methods use the same datasets, their object of investigation is different. It has to be kept in mind that lexicostatistics is concerned with subgrouping whereas glottochronology provides a divergence date to a pair of languages. The merits and demerits of the quantification of time depth in historical linguistics is addressed in a collection of articles edited by Renfrew, McMahon and Trask (2000). The main criticism against glottochronology is that the method works with a constant rate of lexical replacement (in general, language change). However, the recent phylogenetic techniques (cf. section 2.4.2) do not necessarily assume a constant rate of language change. Hence, the trees inferred from modern methods can be dated using sophisticated statistical techniques (Gray and Atkinson 2003). Even McMahon and McMahon (2005), who employ the latest computational techniques from computational biology to classify languages from the Andes and Indo-Aryan languages (McMahon and McMahon 2007) spoken in Northern India, refrain from assigning dates to splits (McMahon and McMahon 2005: 177).

Given the criticism against the (above) presented techniques, how come there are so many posited families? Is the comparative method highly successful in positing these families? In general, the answer is *no*. There are only few language families which are posited by the comparative method. For instance, consider the languages spoken in New Guinea. There are more than 800 languages spoken in the island which do not belong to the Austronesian language family. How are these languages classified? In fact, the recent textbook of Hock and Joseph 2009: 445–454 does not list any of New Guinea's languages. Many of the proposed language families in New Guinea are proposed based on loose cognate counts, similarities in pronouns, typological similarity or geographical proximity (Wichmann 2013, citing Foley 1986).

Long-distance genetic proposals is a contentious topic in historical linguistics. Probabilistic testing of suspected long-distance relationships or linguistic hypotheses is met with skepticism. In a survey, Kessler (2008: 829; my emphasis) makes the following observation:

> Probabilistic analysis and the language modeling it entails are worthy topics of research, but linguists have *rightfully been wary* of claims of language relatedness that are based primarily on probabilities. If nothing else, skepticism is aroused when one is informed that a potential long-

> range relationship whose validity is unclear to experts *suddenly becomes a trillion-to-one* sure bet when a few equations are brought to bear on the task.

Examples of such probabilistic support from (Kessler 2008: 828):

- Nichols (1996) demonstrates that any language with an Indo-European gender system would be, in fact, Indo-European. She did this by counting frequencies of languages that have genders, that a language should have at least three genders, that one of the gender markers should be *-s*, and so on from a large number of languages. The final number for chance similarity is $.57 \times 10^{-6}$ which is such a small number that the original hypothesis cannot be ruled out as a case of chance similarity.
- Dolgopolsky (1986) found similarities between words for 13 concepts and ruled out the chance similarity with a numerical support of 10^{-20}. The small number provides support for a broad Sibero-European language family.

Summarizing, any attempt at comparing the proto-languages of even spatially proximal families is usually viewed with skepticism. The next subsection discusses the reality of linguistic reconstruction and attempts at correlating the linguistic evidence with archaeological and other kinds of evidence.

2.8 Linguistic diversity and prehistory

Linguistic diversity (cf. section 2.1) provides a quantification of the geographical and (indirectly) temporal spread of languages. Explaining the when, where, and why of linguistic diversity has been gaining the attention of scholars in the last few years. These questions have been addressed from different perspectives such as biology, language contact, ecology, and socio-linguistics.

2.8.1 Diversity from a non-historical linguistics perspective

Gavin et al. 2013 is a recent survey of various efforts to explain the extant linguistic diversity in the world. The authors also outline a methodological and theoretical program for future research on linguistic diversity. The paper describes some caveats when employing correlation or regression studies in

studying the causes behind linguistic diversity. Then, the authors propose the following four different processes as drivers of linguistic diversity.

- *Neutral change*: This change is analogous to vertical transmission change in biology where hereditary units, genes, are transmitted from parents to children. Children learn new languages from their parents and environment. When a speech community is isolated from its parent community for a sufficiently long amount of time, the isolated community's language undergoes sufficient neutral changes so that it is no longer mutually intelligible with the parent community.

- *Movement and contact*: Although movement and contact are different factors, they are strongly interrelated drivers of language change. Physical and social barriers are also drivers of language diversification. As such, movement of population groups over long geographic distances followed by subsequent isolation caused languages to diversify. One example is the diversification of Sanskrit into middle Indo-Aryan languages that caused the western Indo-Aryan languages to become mutually unintelligible with eastern Indo-Aryan languages. Population movement can also replace the previously existing languages when entire populations shift from one language to another (usually living some trace of the shift). Contact can also result in language change through the process of borrowing parts of vocabulary or syntax from one language to another; or can result in the development of new languages (creolization).

- *Selection*: This driver of linguistic diversification has two senses. In one sense, it is the social factor that drives the leveling of linguistic varieties in a bounded mutually intelligible linguistic area. Factors such as age, gender, and social class also drive the linguistic diversification of a single language. In another sense, it is individual choice that can drive linguistic diversification. individual choice can stem out of adoption or borrowing from a prestige language that can lead to upward movement in the social or economic status.

The paper also identifies three key variables – demography, environmental heterogeneity, and time – that influence the above four processes of linguistic diversification.

- Technological innovation such as agriculture can spur the population growth rate that gave rise to complex societies which in turn, replaced or assimilated the hunter-gatherer languages present in those areas.

- Just as physical barriers shape language diversification, the same factor also effects the outcome of the language diversification by accelerating the isolation of a particular speech community.
- Time is the most important factor that allows the above processes to gain sufficient momentum to alter the boundaries of speech communities.

The article ends with a note on *linguistic complexity* – causal factors of linguistic disparity (also called typological diversity by Nettle 1999). The authors admit that socio-linguistic factors play a major role in the generation of linguistic variation. The factors which lead to the development of linguistic complexity are:

1. *Small group size*: Whether small group size causes greater linguistic change or not is debated, but there seems to be not much relation between size or spatial spread of language families to the subsistence type of a particular language family (Hammarström 2010; Bowern 2010).
2. *Low levels of contact*: There are studies which show that both grammatical as well lexical units can be borrowed in contact situations.

There are two other factors – large amounts of shared information, high stability – which are not discussed at all.

2.8.2 Diversity from a historical linguistics perspective

Historical linguists posit a reconstructed language (proto-language) through the application of the comparative method. The reality of those reconstructions are always questionable and need to fit with the stories told by other prehistoric disciplines such as genes and archaeology. Heggarty poses the following high-level questions regarding the reality of the reconstructed language families and the corresponding proto-languages.

- When was the proto-language spoken? This question refers to the age or time-depth of the proto-language. An extension would be the time depth of various splits in the proto-language's descendants.
  - Language divergence: This technique is based on correlating the amount of divergence in a language group with its documented time-depth. This approach is at least in two forms. In one form, Holman et al. (2011) employs a regression model to predict the time depths of new families based on linguistic divergence.

Another paper of Bouckaert et al. (2012) employs Bayesian phylogenetic techniques to date the proto-Indo-European to 9,000 years BP. In the case of Austro-Asiatic, Holman et al. give a date of 3600 years which is close to the beginning of the rice cultivation period in the Mekong river delta – 4300 BP.[12]

  - Cultural reconstruction: This technique attempts to correlate the reconstructed proto-forms with technological or material advances recorded in archaeology. The presence of cultivation of specialized forms of agriculture or pastoralist related products is supposed to be interpreted as part of the culture of the proto-speakers.

- Where was the proto-language spoken? This question refers to the putative homeland of the language family.

  - Linguistic paleontology has played a major role in narrowing down the plausible homelands of language families. The Austro-Asiatic language family is a classic example where the reconstructed flora and fauna were used to propose a tropical, coastal homeland and reject any homeland in Southern China (Sidwell 2010).
  - The center of diversity principle is another guiding principle in locating the homeland for a language family. The principle originates with Sapir (1916), who proposed that the geographical location with the largest number of deep branches is supposed to be the homeland. Wichmann, Müller and Velupillai (2010) used lexical divergence as an index to triangulate the homelands of major language families of the world. The focus of diversity principle was also employed, for instance, by the Austro-Asiatic scholars to propose that the Mekong river axis served as the original homeland for the family.
  - Phylogeography is a technique which attempts to situate the language family's phylogeny into the language areas so that a population movement model can explain conflicting hypotheses (Bouckaert et al. 2012).

- Why did the family come into existence? A partial answer to this question has been given by Gavin et al. (2013) in terms of environment, subsistence patterns, conquest, and prestige.

[12] `http://icaal.org/abstract/sidwell-family.pdf`

2.9 Genetics and linguistic prehistory

The recent advances in genetics and comparative linguistics have allowed researchers to make bold claims about the correlation between linguistic history and the speakers' history. The tools developed in genetic studies allow researchers to test specific hypotheses about population prehistory and the prehistory of the languages spoken by those populations.

One such original study is the study of Cavalli-Sforza et al. 1988, who attempt to combine evidence from genes, language, and archaeology to develop hypotheses concerning the history of modern humans. There is also another strand of work by Bolhuis et al. (2014) – who mainly work with generative linguistics – claiming that there is an abstract language organ present in the human mind that allows infants to learn the language from their environment at an astonishingly rapid rate. A third strand of work summarized by Heggarty (2014) attempts to look into the prehistory of a language population(s) through the evidence obtained from archaeology.

Akin to the idea of modern human migrations from Africa (Vigilant et al. 1991; Stoneking 2006) there is the idea of a putative homeland for reconstructed proto-languages in historical linguistics. An example of one such study concerns the putative homeland of the Austro-Asiatic language family (Sidwell and Blench 2011).

Some research consists of approaching language origins from genetic evidence. This research consists of using the accumulated findings from genetics such as Neandertal intermixing with humans and looking at the Hominin anatomical findings to make conclusions about the origins of human language itself. This research consists of looking at the Hominin phylogeny and employing the most parsimonious explanation for human language origin (Dediu and Levinson 2013, 2014).

2.9.1 Early studies and problems

There are at least two early studies that combine genetic history with that of linguistic and archaeological history: Sokal 1988 and Cavalli-Sforza et al. 1988. Sokal (1988) followed a relatively simple procedure when compared to the recent Bayesian methods employed for phylogenetic inference (Ronquist and Huelsenbeck 2003). Sokal compared three different distances – genetic, geographic, and linguistic – in the form of matrix correlations. The individual datapoints were clustered beforehand so that the comparisons are made within systems. Genetic distance is computed as the the mean absolute difference between two allele frequencies. Spatial distances are computed as the

great-circle distances. Linguistic distances follow the disputed classification of Ruhlen 1991.[13]

Sokal performed all possible correlations between the distance matrices. Partial correlations are conducted to negate the effect of one factor while testing the correlation between the other two.[14] Sokal noticed that there are strong partial correlations between genetic distance and linguistic distance as well as geographic distance – if the other distance is kept constant. Sokal observed that more systems show stronger and significant genetic–geographical partial correlations than genetic–language partial correlations. Overall, the paper concludes by pointing out that there is correlation between language and genes even after accounting for geographical factor.

2.9.2 Genetic studies: A world-wide scenario

Pakendorf (2014) reviews three different angles of prehistory investigation that forms the recent interdisciplinary field:

- Coevolution of language and genes.
- Prehistoric contact and its effect on language change and evolution.
- Demographic history of language families and its effect on the prehistory of the language speakers.

The first angle of investigation is concerned with whether linguistic boundaries also act as barriers to genetic flow (*admixture*). There are two models in this investigation: the *branch-split model* and the *isolation-by-distance model*. The branch-split model hypotheses that languages and genes coevolve in a successive split fashion and subsequent isolation. The isolation-by-distance hypothesis is characterized by an inverse relation between increasing linguistic distance and decreasing genetic affinity (Cavalli-Sforza et al. 1988; Sokal 1988). The past studies, undertaken from this angle, suffer from misinformed linguistic classifications.

The second angle of investigation focuses on the undocumented prehistoric contact between two different language groups which can be revealed by genetic analysis. In contrast to the macro-scale correlation studies, this line of investigation focuses on language group specific

[13]Traditional language families are called phyla and subgroups are labeled as language families.

[14]Significance testing is performed by means of Mantel test.

hypotheses concerning the extant spatial distribution of language speakers. Identifying the reasons behind the spread of click consonants in Bantu languages is an example of such a study. The reason for the borrowing is attributed to maternal ancestors (female speakers) belonging to click consonant language groups (Bostoen and Sands 2012).

The third angle of investigation uses language family histories to infer the language family's prehistory. This is done by analyzing the lexical cognate data for members of a language family and inferring a phylogenetic tree through a off-the-shelf phylogenetic software and test the different possible hypotheses of homelands. Some examples of such studies for different language families and geographical areas are:

- Indo-European: Bouckaert et al. (2012), Gray and Atkinson (2003).
- Austronesian: Gray and Jordan (2000), Jordan et al. (2009).
- Bantu: Holden (2002), Holden and Gray (2006).
- Tupian: Walker et al. (2012).
- Dene-Yeniseian: Sicoli and Holton (2014) (also uses structural features).
- Melanesia: Hunley et al. (2008).
- Southeast Asia: Donohue and Denham (2011), Denham and Donohue (2012).

There are also studies which employ structural features for inferring phylogenies such as Dunn, Levinson and Lindström (2008) who attempt to distinguish Papuan languages from Oceanic languages.

Finally, there are some genetic studies which were conducted on a geographical micro-scale for language groups from rest of the world.

- Indigenous American languages: Reich et al. (2012), Amorim et al. (2013) (Only for South America)
- Indo-European: Balanovsky, Utevska and Balanovska (2013) and Burlak (2014).
- Bantu languages: Quintana-Murci et al. (2008).

In conclusion, a combination of computational, statistical, linguistic, and anthropological techniques can help address some questions about the origin and spread of language families both spatially and temporally.

2.10 Conclusion

This chapter presented a linguistic introduction to the processes of linguistic change, models of language evolution, computational modeling of the linguistic changes, and the recent developments in computational historical linguistics. The next chapter will summarize the various linguistic databases that resulted from digitization as well as new efforts to augment the older vocabulary and typological databases.

3 DATABASES

This chapter describes the various linguistic databases used for language classification. The papers listed in the third part of the thesis describe the Automated Similarity Judgment Program (ASJP) database and World Atlas of Language Structures (WALS) database. Thus, this chapter will focus on linguistic databases which are not listed in part III of the thesis. The linguistic databases used in language classification can be classified into the following three types.

- *Cognate databases.* Linguistic databases that show the cognacy status of phonological, lexical, and grammatical features (characters) across a language family.
- Typological databases presenting the variation of a typological feature on a graded scale.
- There are other linguistic databases that show linguistic features such as phoneme inventory size and part-of-speech annotation.

3.1 Cognate databases

Core vocabulary databases are parallel word lists for a language group. The size of the word lists usually range from 40 to 215 in these databases. The basic vocabulary databases are lexical in nature and may also carry cognate judgments. The core vocabulary databases can be used for lexicostatistical studies and also as an input to the distance-based or character-based phylogenetic algorithms (cf. section 2.4.2).

3.1.1 Dyen's Indo-European database

Dyen, Kruskal and Black (1992) prepared a lexicostatistical database of 95 Indo-European speech varieties for 200 concepts. The database has word

forms and cognate judgments for the Celtic, Germanic, Indo-Iranian, Baltic, Slavic, Greek, Armenian, and Albanian branches of IE. The word forms in the database are not phonetically transcribed and hence, are not fit for phonetic analysis or computing phonetic similarity distances between the speech varieties. However, the database was used for the purposes of cognate identification and inference of a Levenshtein-distance based IE tree (Ellison and Kirby 2006).

3.1.2 Ancient Indo-European database

Ringe, Warnow and Taylor (2002) designed a database consisting of IE word lists for 24 ancient Indo-European languages. The database has 120 concepts in addition to the 200 Swadesh concepts, 15 morphological characters, and 22 phonological characters. Each character can exhibit multiple states. The presence of the *ruki* rule – change of PIE */s/ to */š/ after */r/, */u/, */k/, or */i/ – is coded as 2 in Albanian, Armenian, Indo-Iranian, and Balto-Slavic languages and its absence as 1 in other IE languages. Whenever a meaning has two forms, each form is coded as a separate character and the cognate judgments are assigned accordingly. For instance, Luvian shows two word forms for the concept 'all (plural)'. Each word form is cognate with word forms present in some other IE languages. Thus, the two word forms are listed as separated characters. Nakhleh et al. (2005) compare the performance of various distance-based and character-based algorithms on this dataset.

3.1.3 Indo-European Lexical Database

IELex (Bouckaert et al. 2012) is a lexical database of Indo-European doculects (135)[15] compiled by Michael Dunn. IELex is compiled from Dyen's Indo-European database and Ringe's ancient IE languages database. The cognate judgments from both the databases are refined and included in IELex database. As of now, the lexical forms' transcription is in both orthographic and IPA format.

3.1.4 List's database

List and Moran (2013) developed an python-based open-source toolkit for CHL. This toolkit implements the pipeline described in chapter 2 (cf. figure

[15] Accessed on 15th September 2015. The database is available at `http://ielex.mpi.nl/`.

2.6). The authors also provide a manually curated 200-word Swadesh list for the Germanic and Uralic families, Japanese and Chinese dialects. The word lists are encoded in IPA and the toolkit provides libraries for automatic conversion from IPA to coarser phonetic representations such as ASJP and Dolgopolsky's sound classes.

3.1.5 Austronesian Basic Vocabulary Database (ABVD)

ABVD[16] (Greenhill, Blust and Gray 2008) is a vocabulary database for 998 Austronesian languages. The database has 203,845 lexical items for the Swadesh concept list (of length 210). The database has cognate judgments and has been widely used for addressing a wide-range of problems in Austronesian historical linguistics (Greenhill and Gray 2009).

3.2 Typological databases

3.2.1 Syntactic Structures of the World's Languages

Syntactic Structures of the World's Languages (SSWL)[17] is a collaborative, typological database of syntactic structures for 214 languages. Although the data is available for download, not much is known about the current state of its development.

3.2.2 Jazyki Mira

Jazyki Mira is a typological database which is very much like WALS but with fuller coverage for a smaller set of Eurasian languages (Polyakov et al. 2009). Polyakov et al. (2009) compare the calculations of typological similarity and temporal stability of language features from the data obtained from WALS and Jazyki Mira.

3.2.3 AUTOTYP

AUTOTYP (Autotypology) is another typological database based at the University of Zurich (Bickel 2002). Rather than working with pre-defined list of typological features, the project modifies the list of typological features as

[16] Accessed on 2nd December 2013.

[17] `http://sswl.railsplayground.net/`

more languages enter into the database. The database was used for investigating quantitative and qualitative typological universals (Bickel and Nichols 2002).

3.3 Other comparative linguistic databases

There are some databases which are indirectly related to CHL but so far have not been employed for language classification.

3.3.1 Intercontinental Dictionary Series (IDS)

IDS is an international collaborative lexical database for non-prestigious and less known languages. The database is organized into 23 sections consisting of 1,310 concepts. The database has a large collection of languages from South America and the Caucasus region. The database has 215 word lists which are available for online browsing and download (Borin, Comrie and Saxena 2013). An extended concept list is proposed in the *Loanword Typology Project* (LWT) described in the next section. Cysouw and Jung (2007) use the IDS word lists from English, French, and Hunzib for cognate identification through multi-gram alignments.[18]

3.3.2 ODIN

Online Database of Interlinear Text (ODIN; Lewis and Xia 2010) is an automatically extracted database from scholarly documents present on the web. The ODIN database has more than 190,000 instances of interlinear text for more than 1,000 languages. The authors parse the English gloss text and project the syntactic structures to the original language data creating a parallel treebank in the process. The users can search the database for syntactic trees and categories. The database provides search facilities for searching the language data and the source of the data and is available for download.

3.3.3 World loanword database

The *World Loanword Database*, under the auspices of LWT, is a collaborative database edited by Haspelmath and Tadmor (2009a). This database is an

[18] An n-gram of length i in language A is mapped to an n-gram of length j in language B where $1 \leq i, j \leq n$.

extension of the concept lists proposed in the IDS project. The meanings are organized into 24 semantic fields. For each concept, the database contains word forms, the gloss of a word form, the source of the borrowing (if it is a borrowing) and the expert's confidence on the borrowing on a scale of 1–5, and the age of the word for 41 languages. The age of the word is the time of the earliest attestation or reconstruction for a non-borrowed word; for a borrowed word, age is the time period in which the word was borrowed. Tadmor, Haspelmath and Taylor (2010) apply the criteria such as (a) fewest borrowed counterparts (borrowability), (b) representation (fewest word forms for a meaning in a language), (c) analyzability (for a multi-word expression), (d) age, in order to arrive at a 100-word list called the Leipzig-Jakarta list. The 100-word Leipzig-Jakarta concept list has 60 concepts in common with the 100-word Swadesh list. Holman et al. (2008a) develop a ranking procedure to rank the meanings of the 100-word Swadesh list according to lexical stability and correlate stability ranks and borrowability scores from then still unpublished results of the LWT, finding the absence of a correlation, suggesting, importantly, that borrowability is not a major contributor to lexical stability.

3.3.4 PHOIBLE

PHOnetics Information Base and LExicon (PHOIBLE)[19] is a phonological and typological database for more than 600 languages. The database has phonemic and allophonic inventories, and the conditioning environments that are extracted from secondary sources like grammars and other phonological databases (Moran 2012).

3.3.5 World phonotactic database

The World phonotactic database has been recently published by a group of researchers at the Australian National University (Donohue et al. 2013). The database contains phonotactic information for more than $2,000$ languages, and segmental data for an additional $1,700$ languages. The main focus of this database is on the languages of the Pacific region.

[19] Accessed `http://phoible.org/` on 2nd December 2013.

3.3.6 WOLEX

The *World Lexicon of Corpus* is a database of lexicons extracted from grammars and corpora for 47 languages by Graff et al. (2011). The website[20] lists the 47 languages, size of lexicon, and the source of data. Not much is said about the methodology and development of the corpus on the website of the project.

3.4 Conclusion

In this chapter, various linguistic databases are summarily described. Not all of the databases have been used for language classification. To the best of our knowledge, except for the Ancient languages IE database and ABVD, the rest of the databases have not been exploited to their fullest for comparative linguistic investigations. As noted by Borin, Comrie and Saxena (2013), using larger word lists (such as IDS) would be useful in investigating the *rarer* linguistic phenomena since the data requirements grow on an exponential scale.

[20] `http://qrlg.blogspot.se/p/wolex.html`

Part II

Prolegomenon

4 COGNATES, N-GRAMS, AND TREES

This chapter will discuss the reproduced publications related to string similarities, language classification, lexical similarity, subsequence kernels, and cognate identification. All the examples from ASJP database are given in bold face.

4.1 Some questions

The publications in this thesis work with Swadesh word lists. Since historical linguists work with more evidence than Swadesh word lists, the results obtained from Swadesh lists might be viewed with skepticism by historical linguists. At least the following questions arise when the publications – employing computational techniques – employ 40–200 length word lists for inferring a family tree or make a projection regarding the time depth of splits in language families.

For instance, consider the task of inference of a family's internal structure. Until now, this task is predominantly addressed by the two following computational methods: distance-based (including string distances) and character-based methods (cf. section 2.4.2). The publications reproduced in part III employ distance-based methods. The distance-based methods require raw word lists whereas the character-based methods require predetermined cognacy information. The distance-based methods exploit the observation that cognates in short word lists exhibit more form similarity than the cognate forms for meanings that are less basic. For instance, German **nam3** and Hindi **nam** 'name' are cognates that have three phonemes in common. On the other hand, character-based methods such as Bayesian inference methods employ the notion of loss or gain of cognates along a tree's branches to model language evolution. The inferred tree from the Bayesian methods requires lexical cognate data for each meaning whereas the distance-based methods do not require the cognacy information beforehand to infer a tree.

Publication VI (chapter 12) on the comparison of string similarity measures works with raw word lists where the cognacy information is not available beforehand. In fact, many of the world's language families do not have sufficient scholarship to determine the status of cognacy even for short word lists. Also, distance-based methods are computationally cheap and fast when compared to the Bayesian methods. The question that arises is if the distance-based methods are reliable? The following questions can be raised with respect to the distance-based methods:

1. What linguistic features are important for identifying the probably related languages?
2. How to automatically identify cognates from the word lists?
3. Is it justified to go from raw word lists to family trees?

4.2 Linguistic features for probably related languages

The first question corresponds to the methodological issue raised by Nichols (1996) when beginning to apply the comparative method to a group of probably related languages. The issue brought up by Nichols suggests that linguists employ some notion of similarity to group different languages into a language family. This similarity can be in terms of shared phonology, highly similar lexical or grammatical items. As a computational linguist, one task would be to develop a reliable computational technique that can distinguish related languages from unrelated languages. At this stage, the loanwords and chance similarities are not identified. Note that the character-based methods will not work at this stage due to the lack of cognacy information.

One way to quantify the phonological similarity between two words is to use Jaccard's index that is defined as the number of shared phonemes divided by the total number of phonemes. If two words are related by descent then they might share some phonemes depending on the extent of phonological change that happened in both the languages after the divergence from their ancestral language. If one language has undergone lexical replacement or a large phonological change such that there are no shared phonemes between two words; then Jaccard's index would yield a similarity of zero in such a case.

4.3 Cognate identification and language classification

The second question is related to the problem of cognate identification. Given the word lists for a group of related languages, how can the cognates be identified between the related languages? A number of string similarity measures can be applied to quantify the similarity or dissimilarity between two words if the words are treated as a sequence of characters/phonemes/segments.[21] There are a number of string similarity measures in computer science (Cohen, Ravikumar and Fienberg 2003) that exploit different properties to quantify the similarity between two words. Levenshtein distance or edit distance, introduced in chapter 2, is one such measure. The edit distance score between two words can be converted into a similarity score by subtracting the distance score from one. The edit distance also yields a character alignment between the two strings. However, edit distance is coarsely defined and can be modified to take the systematic sound correspondences that occur in related languages.

The similarity counterpart to edit distance is longest common subsequence ratio (LCSR). Subsequence formulation of string decomposition allows one to define sequences that are extracted by dropping intermittent characters in a given string. For instance, dropping the *o* from English *hound* yields Norwegian Riksmål *hund*. Note that by using Swadesh lists for cognate identification, we are attempting to identify cognates among those lexical items that have not undergone substantial semantic change. The string similarity methods can be applied to identifying grammatical morphemes also.

Both LCSR and edit distance belong to the family of dynamic programming algorithms where one can trace the working of the algorithm by filling a chart when computing the similarity or distance between two strings. Edit distance, as defined originally, works with a single character and attempts to match those characters that are completely similar to each other. For instance, edit distance would give the edit distance between *pin* and *bin* as 1 since both the words differ in the first position. The distance between *pin* and *zin* would be the same since it does not distinguish the difference in similarity between [p] and [b] as compared to [p] and [z].

The graded notion of similarity (or difference) between two phonemes is dependent on both phonology and phonetics. If the strings are transcribed in IPA, then the difference between two strings can be quantified in terms of pre-defined weights. The difference between [p] and [b] would be voicing. The

[21] In this section, character means element of a string.

similarity between [p] and [b] would be the weighted similarity in place and manner.

There is also an empirical component involved when linguists work with sound correspondences. The character alignment produced by edit distance can be used to produce a frequency table of character correspondences with counts. These counts can be employed to assign an importance to the sound correspondence. For example, a sound correspondence such /s/ $\sim$ /h/ is quite common across the world's languages (Brown, Holman and Wichmann 2013). Note that the frequency of the individual segments, /s/ and /h/ is also high across languages. This is again in line with the observation that some phonemes occur across the languages whereas some phonemes are distinctive for a language family. For instance, retroflex consonants are generally a distinctive feature of Dravidian and Australian languages.

The idea of sound correspondences is in parallel with that of collocations in computational linguistics. The importance of a lexical collocation is estimated by the use of PMI (point-wise mutual information; Church and Hanks 1990) which is defined as followed:

$$pmi(x;y) = log\frac{p(x,y)}{p(x)p(y)} \tag{3}$$

The above definition exploits the information about the frequency of segments in the word lists. If a correspondence is rare but its segments are highly frequent then the correspondence has low information content. A frequent correspondence will also have low information content since the segments are also highly frequent. However, if the segments are not of high frequency but the correspondence is a frequent correspondence, then the correspondence has a high information value. Such high information correspondences are important for cognate identification. The PMI score is also a kind of statistical test that tests if the sound correspondence occurs by chance.

Since we treat each translation pair as a potential cognate, the above procedure is iterated through few times until the values of the PMI score do not change significantly. Since the set of phonemes are fixed beforehand, the PMI matrix consists of cells filled by the PMI scores between each segment pair. One can expect positive values for vowels against vowels and consonants against consonants. For example, Jäger (2013) obtains a positive score of 0.36 between [p] and [b] as compared to -7.07 between [a] and [b]. It has to be noted that Wieling, Prokić and Nerbonne (2009) developed this scoring procedure for dialect data that is neatly organized into cognate sets and the authors did not have to work with the notion of potential cognates.

The similarity scores between two phoneme segments are then employed to compute the similarity between two words. There are some variations as to how to compute the similarity between two words. The algorithm used to compute the longest common subsequence can also be employed to compute the character alignment(s) with the maximum similarity (Needleman and Wunsch 1970). Edit distance is the distance counterpart to this similarity algorithm. Wieling, Prokić and Nerbonne (2009) convert the pair-wise segment similarity matrix to a distance matrix and then employ it to compute the minimal distance alignment between two strings. The authors use the standard normalization technique to convert a PMI score, which is negative or positive, into a score between 0 and 1, such that the lowest PMI score gets a distance of 1 and the highest PMI score gets a distance of 0. Kondrak (2002a) also uses the similarity algorithm to align a pair of strings. Maximizing the similarity and minimizing the distance have been shown to be equivalent (Sellers 1974).

Downey et al. (2008) propose a new way of converting a string similarity score between two strings S_a, S_b into a distance score as follows:

1. Compute the similarity of a string S_a from itself. Do the same for S_b also.
2. Divide twice the similarity between S_a, S_b by the sum of the self-similarities. This division converts the similarity into a score less than 1.
3. Subtract the similarity score from 1 to convert it into a distance.

The distance score is 0 for identical words and less than 1 for non-identical words. The string alignment techniques will not be able to detect cognates that underwent drastic shifts such as Armenian *erku* $\sim$ English *two* 'two' (Hock and Joseph 2009: 583–584) given earlier. To the best of our knowledge, there has been no study that compares different techniques to convert similarity to distance scores for the task of language classification or cognate identification.

Until now most authors employed single character matches for identifying the similarity between two words. Contiguous slices of characters – n-grams – with length greater than 1 can also be matched. For example, the similarity between *hound* and *hund* can be computed in terms of the common bigrams shared between the two words. This kind of similarity brings in some context when comparing two words. The words differ in terms of *ho, ou*. This kind of similarity is trivial to compute since one only has to extract the n-grams in each word and enumerate the different n-grams. The bigram similarity notion can be extended to trigrams and to higher order of grams. Figure 7.4 shows that the mean of mean word length in many of world's language families is around 4.

So the inclusion of 4-grams or 5-grams would include the whole word into the comparison.

The problem of segment alignment has a connection to the problem of letter to phoneme conversion (L2P), machine transliteration task in NLP, and word alignment in statistical machine translation (SMT). For instance, the problem of word alignment is a well-studied one in SMT (Och and Ney 2003). SMT employs parallel corpora of large size – a few million sentences – for aligning words from a source language to a target language. A typical SMT word alignment program such as the Berkeley word aligner (Liang, Taskar and Klein 2006) outputs a mapping between each sentence pair. The word alignments are not known beforehand and are learned automatically over a number of iterations until an objective function is minimized. In contrast, the number of units of alignment in L2P or transliteration task is quite small (below 100). This allows one to investigate if *off-the-shelf* SMT systems such as Moses (Koehn et al. 2007) can be used to align two words. One driving idea behind the use of sequence alignment data is the large human effort required to build cognate databases like ABVD (Greenhill, Blust and Gray 2008) for the Austronesian language family and the Indo-European database (Dyen, Kruskal and Black 1992); and even align the words manually (List and Prokić 2014).

An effort by Kondrak (2009b) consists in employing word alignment models to align segments in words. As noted above, the string alignment algorithms yield a one-to-one segment correspondence between words whereas the SMT techniques can be used to extract complex segment correspondences, also known as many-to-many word alignments in SMT literature. One example of such complex segment correspondences occur in the Algonquian languages (Campbell 2004: 141) and is reconstructed for the Proto-Algonquian language. Rama, Kolachina and Kolachina (2013) used the segment alignments extracted from the MOSES phrase alignment system to compute the PMI-based similarity score between two multiple length segments and then use the PMI-based similarity score to compute the edit distance between two words. The results suggest that the method yields comparable results with PMI-based weighted edit distance. This method has an advantage that it does not require manually annotated correspondences between two words and can be employed to yield a list of tentative sound correspondences between those languages that have not been so well-studied.

The string similarities comparison paper (cf. publication VI) does not perform explicit cognate identification and employs raw string similarity measures to separate related from unrelated languages. The main goal behind the work presented in the publication is to perform a comparative evaluation of different string similarity measures for automatic language classification.

As noted by Jäger (2013), the ultimate goal of different methods is to perform language classification automatically. At this point, the discussion can branch into two paths. One path works with explicit cognate identification and its subsequent use for automatic language classification. The other path works with raw word lists for automatic language classification.

4.3.1 N-grams for historical linguistics

In computational linguistics, character level n-grams have been employed for various tasks such as language identification (Cavnar and Trenkle 1994), document visualization (Jankowska, Keselj and Milios 2012), authorship attribution (Escalante, Solorio and Montes-y Gómez 2011), machine translation for closely related languages (Nakov and Tiedemann 2012), and named-entity recognition (Klein et al. 2003).

How can one relate n-grams to processes of sound change in historical linguistics? A reflex can undergo phonological change to represent some of the changes that happened since the language descended from its common ancestor.

An extended definition of n-grams called *skip-grams* captures the loss or gain of phonemes in a daughter language. There are some string similarity measures that use skip-grams to compute the similarity between two strings. For example, XDICE measure is a variant of DICE measure that uses skip-one bigrams. An example of such skip-one bigrams in *hound* is *hu, on, ud* and *hund* has *hn, ud*. In this example, one character is skipped to form skipped bigrams. Apart from the regular bigram similarity, both strings share an extra skipped bigram of *ud*. In principle, the definition of skip-grams allows multiple skips to form skip-grams. Skip-grams are also known by the name of subsequences in machine learning literature. The subsequences are studied under the title of "string kernels" in this literature. Skip-grams can capture long-range effects in words. Sound changes such as assimilation and dissimilation can be captured by skip-grams. The property to skip allows a string similarity definition that can be employed for cognate identification. Note that the notion of skip is analogous to the notion of *gap* or deletion in edit distance.

I will illustrate the utility of skip-grams in the case where one form appears to be different due to intermittent segments but the other form has not undergone such change. Consider the German $\sim$ Swedish word-pair for 'star', *StErn* $\sim$ *SEnE*, from ASJP database. The Swedish form lost [t]. The words do not have any common bigrams but share two skip-one bigrams *SE, En*.

A characteristic property of a language family is its syllable structure. This property is usually expressed in terms of C/V patterns. This property can also

be used for cognate identification. In publication VII (chapter 13), I use this property to compute the weighted similarity between two words.

Kondrak (2009b) is one study that employs word alignment models to learn many-to-many segment correspondences in Algonquian languages. Some examples of such correspondences from Fox–Menomini language pair are **hk** ∼ **hk**, **ht** ∼ **qt**, **t** ∼ **ht**. Another example of such many-to-many correspondences is Swedish Romani *ra**tt**i* ∼ Hindi *ra**tr**i* 'night' where **tt** ∼ **tr** show the assimilation that took place in Swedish Romani.

In an earlier paper, Kondrak (2005) introduces n-gram similarity to extend the edit distance formulation that works with unigrams. Kondrak's definition of n-gram similarity is more general such that it can capture three different kinds of similarities between the bigram sets of both the words. The bigram set for *ratti* is *ra, at, tt, ti* and for *ratri* is *ra, at, tr, ti*. The three similarities are defined as followed:

1. Binary similarity: Two bigrams are similar if they are completely similar. This means that *tt* and *tr* are not similar.

2. Position-wise similarity: It is well known in historical linguistics that some sound changes are dependent on the position of the sound in the word. *tt* and *tr* have a similarity of 1 since they have same unigram *t* in the first position and different unigrams in the second position.

3. Comprehensive similarity: Compute the number of common unigrams between two bigrams. Comprehensive similarity is useful when there are cases of metathesis which is not captured by the position-wise similarity.

The position-wise and comprehensive similarities are soft matching similarities. The bigram similarity definition captures the context when comparing segments. One example of such complex segment correspondence is English /p/ ∼ German /pf/ in the word-initial position and English /p/ ∼ German /ff/ elsewhere.

In publication VII (chapter 13), both character subsequences and CV string subsequences are employed to compute word similarity. One intuition behind the idea that skip-a-character(s) will work is the observation that consonants are generally more stable and vowels change more frequently. So, one way to quantify word similarity would be to look at the consonant similarity. For example, *hound* and *hund* have the common consonant skeleton *hnd*. The CV string for the word pair is *CVVCC* and *CVCC*. The skip-grams of the CV strings are also compared in order to compute a word similarity score. The results presented in publication VII (chapter 13) work for skip-grams of length 2 on the task of cognate identification. When two

words do not show any phonological similarity and are nevertheless cognates, converting the words from IPA to coarse-grained transcription such as ASJP alphabet or Dolgopolsky's sound class string (Dolgopolsky 1986) can help in identifying the similarity between the two words.

The connection between the topic of language identification and language discrimination can be motivated as follows. The language identification task, in NLP, is about the automatic identification of languages in multilingual texts. The main idea behind the language identification approach is that widely different languages have different n-grams as features that can be employed to train classifiers for classifying multilingual texts. Porta and Sancho (2014) observe that character n-grams work as good features for classifying unrelated languages but do not work so well for discriminating closely related languages or varieties (such as Hindi and Urdu). Does this hold for a language family also? Do language families differ in terms of the high frequent character n-grams that make up the words in the family's language members? This question can be answered by looking at the frequency of phonemes in language families' word lists in ASJP database. For example, the dominating segment in Khoisan languages is the click sound; in Dravidian, the single retroflex consonant[22] is a highly frequent segment in the n-grams; the Niger-Congo languages show a high frequency of nasal vowels; and Mayan languages show a high frequency of a glottal modifier **'** that represent the glottal sounds. The high frequency of such signature phonemes suggests that the n-gram features can be used to separate a language family from all the other families.

One intuition about the n-grams is that they can also serve as a net index of the amount of language change that happened in a language family. The language change is typically lexical replacement and phonological change. The change of status in sounds from an ancestral language to its daughter languages can be summed by looking at the types of attested n-grams in the daughter languages. This idea is utilized for linking the time-depths of a family to its n-gram diversity (cf. chapter 5) which is discussed in publication IV.

4.3.2 Language classification with cognate identification

The first path of automatic cognate identification requires manually annotated data which is split into training and testing parts for training a automatic classifier. To evaluate any automatic cognate identification system, the first hurdle is the availability of human annotated gold standard data. Only Indo-European and Austronesian languages have such kind of data for

[22] A single symbol is used to represent all rhotic sounds.

200-meaning Swadesh lists. It has to be noted that both families have more than a hundred years of scholarship to arrive at this stage.

Jäger (2013) circumvents the gold standard data paucity problem by designing an objective function that attempts to tie in language relatedness and cognateness in the same function. Kondrak (2002a) does not use a single threshold to determine the cognacy status but uses a evaluation score called *11-point interpolated precision* to evaluate his system's performance. Rama, Kolachina and Kolachina (2013) employ the PMI-based technique to determine the similarity between two phoneme segments. Their method does not require any training data. Their method stops training when two subsequent iterations do not show any significant improvement between the PMI similarity matrices. Finally, they use Pearson's r to evaluate their system's performance against the binary gold standard data.

The driving idea behind the use of string similarity measures is that some cognates are identical as compared to other cognates that are not at all similar. For instance, Armenian *erku* and English *two* are cognates that can be traced back to the Proto-Indo-European language. However, the automatic methods will not be able to find such cognates which have changed to a great extent. At this point, it is worth mentioning the lack of graded notion of cognacy. Historical linguistics normally assign a binary status of cognacy vs. non-cognacy to a word pair. Some cognates such as Swedish **tv~o** and Hindi **do** 'two' are easy to identify due to the regular sound correspondence between the initial segments of the two words. On the other hand, the Armenian-English cognate pair *erku* $\sim$ *two* 'two' shows that such cognates are tough to identify. Authors like Campbell and Poser (2008: 173) characterize such cognates as "true but non-obvious". Two more examples of such non-obvious cognates from Campbell and Poser 2008: 173 are Armenian *hing* $\sim$ English *five* and French *cinq* $\sim$ Russian $p^{j}at^{j}$ 'five'.

At this stage, we have a a weighted string alignment method that computes the similarity or distance between two strings based on empirically determined alignments. The relations between a set of languages can be computed either through aggregation or by clustering the pair-wise language matrix for each meaning to obtain cognates.

- The aggregation method consists of summing the lexical distances between each meaning and averaging it to obtain a net distance score. The pair-wise distance matrix is then supplied to a tree-inference algorithm to infer a family tree.

- For a meaning, the string alignment method provides a similarity score between each pair of languages. The pair-wise similarity scores are supplied to a clustering algorithm. The clustering algorithm typically

yields a tree for the distance matrix. This step is applied to each meaning's distance matrix to infer a tree. The trees for each meaning can then be combined to infer a consensual tree to show the relations between languages.

The second way of aggregating the distance relations follows the dialectologists' maxim that "each word has its own history". Each tree for a meaning is unique and represents a history of the meaning.

4.3.3 Language classification without cognate identification

This subsection attempts to answer the third question listed in the first section. Is it justified when an attempt is made to infer family trees without cognate identification? There are two steps involved in this discussion: language discrimination and classification. The discrimination step is evaluated by the use of distinctiveness. The classification step is evaluated by comparing the inter-language distance matrices to the tree classification taken from Ethnologue.

In publication VI (chapter 12), we does not perform explicit cognate identification for automatic language classification step. In this publication, we attempt to find which measure is the best task for language discrimination. The ranking in the publication shows that Jaccard's index followed by trigram and bigram similarity measures perform better than edit distance for language discrimination.

The work of Kessler (2001, 2007) employs the *permutation test* for the purpose of testing the language relatedness hypotheses. Kessler (2007) tests the hypothesis if some Uralic languages and Indo-European languages are genetically related by performing a permutation test. The permutation test, for a pair of languages, is described as followed:

1. Compute a word similarity score between all the lexical items that belong to the same meaning.
2. Do the above step for all the meanings and sum it up. This number shows the net word similarity between languages assuming that they are related.
3. Permute the items randomly within a language so that the original meaning-form setup is disturbed. Do the above two steps with the permuted items.
4. Repeat the third step a large number of times, about 10000 times, and average the similarity score.

5. Divide the same-meaning items' similarity score by the permuted items' similarity score. The number of times the numerator is less than or equal to the denominator gives a significance score to the null hypothesis that the languages are not genetically related. If the ratio is less than a predefined value of 0.05 than the null hypothesis is rejected.

The underlying idea behind the permutation test is that if some items are similar due to chance or due to the similarities in phonotactics or phonological inventories such similarities will not show up when items between non-matching meanings are compared. This testing is a way to employ the guiding principle that recurrent sound correspondences are useful for establishing language relatedness. The double normalization introduced by ASJP is a special case of the permutation test where the denominator is the average of all the non-matching meaning items' word similarities. For instance, if there are 40 items each in two lists, then the denominator would have $(40 \times 40) - 40 = 1560$ word comparisons.

Kessler (2001: 154–156) also tests the relevance of some phonetic features for language discrimination. The author uses a simple measure of word matching based on the phonetic feature matching of the first phoneme in a word pair. For instance English **fiS** and Latin **piskis** have a word similarity of 1, based on place, since /p/, /f/ are labial whereas they differ in the place of articulation. The author finds that vowel height, place of articulation, and articulator are significant indicators of relationship between languages.

Kessler (2007) extends the initial phoneme comparison to *sound class* comparison. The sound class approach assumes that phonemes falling within a sound class tend to change to each other more readily than to those phonemes in a different sound class. The ASJP alphabet given in tables 7.1 and 7.2 are an example of sound classes where some distinctions such as vowel length and voicing are collapsed.

Dolgopolsky (1986) also came up with his own ten sound class system that collapses fine distinctions in place and manner of articulation and excludes vowels and this was used by Kessler to compare words. Kessler (2007) uses a definition of word similarity that is based on Dolgopolsky's sound classes. Kessler's word similarity is computed as the number of phonemes that fall into the same sound classes. Kessler uses this measure to test the hypothesis of relation between different language groups such as Indo-European, Indo-Uralic, and Balto-Slavic. His experiments suggest that word similarity measures based on sound classes perform well at identifying established groups as opposed to speculative connections such as Indo-Uralic. This discussion is intended to support the claim that double normalization has a statistical motivation that explains why it performs well at the task of

language discrimination. The distinctiveness score is an index of discrimination which is discussed below.

The distinctiveness score for a family is defined as the difference between the mean of distances within a family to the mean of the distances from the languages in the family to the rest of the languages outside the family. The hypothesis is that the the mean of the intra-family distances should always be smaller than the mean of the distances from the family to rest of the languages, otherwise the languages cannot fall under a single family. The difference is divided by the standard deviation of the rest of the family distances. The standard deviation measures how variant the distances from a family are to the languages falling outside the family. The dist scores given in table 12.2 for each of the string similarity measure merit some discussion as to what they measure. Jaccard's distance shows the highest distinctiveness score across all the families.

The use of bigrams appears to be a better indicator of genetic descent than edit distance. The distinctiveness results are presented in table C.2. Some families that have been established by the comparative method show a higher distinctiveness score than controversial families such as Nilo-Saharan and Niger-Congo. For instance, a small family like Dravidian languages has a high distinctiveness score of 18.5943. The world's largest language family Austronesian shows a distinctiveness of 7.667. New world language families such as Mayan, Mixe-Zoque, and Uto-Aztecan show a distinctiveness score that is greater than the average distinctiveness score computed from all the families. The African language families such as Nilo-Saharan, Niger-Congo, and Afro-Asiatic have distinctiveness scores that are less than 3 and are among the lowest of the scores. Campbell and Poser (2008: 120–144) criticize these language families since they are not established through the comparative method but in terms of typological similarity that might not be convincingly genetic but can arise due to contact also.

One final question that remains to be answered is if it is the double normalization or Jaccard's index that is actually performing so well at the task of language classification. This question can be answered by looking at the significance ranking of the string similarity measures. The best performing measure is JCDD (Jaccard's index's double normalization version) followed by JCD. This ranking suggests that a combination of bigrams and double normalization works the best whereas bigrams also work better than the rest of the measures at discriminating related from unrelated languages. This suggests that the bigrams capture the difference between the languages in a family when compared with the rest of the languages.

Now, I will turn to the second issue of accuracy at language internal classification task. Language trees are a diagram showing how the languages came to exist, as we know, from a reconstructed ancestral language. As such a family tree can be decomposed into triplets – three-language units – that show

the divergence history in a family. Some triplets are not resolved or do not show binary splitting. The reason for the unresolvedness could be due to the following reasons:

- Insufficient research or lack of information about language splitting at such nodes.
- An ancestral language split into multiple daughter languages about the same time. This would also result in a multifurcating node.

The Indo-European tree in figure 2.2 is a multifurcating tree at the top level. If three languages belonging to three different branches of the family are considered – Sanskrit:Indo-Iranian, Latin:Romance, and Ancient Greek – then the family tree shows them to have split from the Proto-Indo-European ancestor at about the same time. The tree does not show which language split first from the proto-language. The Dravidian tree is an example of a multifurcating tree which explains the second scenario that has been listed above. The Dravidian family tree shows a ternary branching at the Proto-Dravidian level where linguistic evidence supports the setting up of such three-way split at the top node (Krishnamurti 2003: 492).

The γ measure (cf. section 12.6.3) is used to compute the agreement between the distance matrix and the gold standard tree is straightforward. The γ measure checks how many subtrees match or mismatch with the gold standard tree. The γ measure is always between -1 and $+1$. The results given in table 12.2 tell that the agreement is always positive. The edit distance's γ score agrees largely with smaller families. For instance, the Mayan family is relatively small and its internal structure is well-resolved. The γ measure is agreeing with 78% of the subtrees in the gold standard. The agreement with Dravidian is also about the same. Dravidian family's gold standard tree is largely bifurcating at different levels except at the top level. The Austronesian family tree is a unbalanced tree with most of the languages falling into Malayo-Polynesian branch and the rest of the languages falling into the nine branches. All the latter branches are situated in Taiwan (Blust 2011). The agreement for the family is consistently low across all the string similarity measures.

The next question in the classification story is: how does one choose the best system? The answer for this comes from the multiple-testing procedure explained in section 12.7. The testing procedure suggests that n-gram similarity measures work best for separating related from unrelated languages whereas there is no significant difference across string similarity measures at the task of internal classification.

5 LINKING TIME-DEPTH TO PHONOTACTIC DIVERSITY

In this chapter, I will discuss the statistical model employed in publication IV for predicting the time depth of language families across the world. In the publication IV, I employ phoneme n-gram type diversity as a predictor of the time depth. The intuition that n-grams are an index of time depth comes from observation made by Sapir (1916):

> The greater the degree of linguistic differentiation within a stock, the greater is the period of time that must be assumed for the development of such differentiations.

The above observation suggests that if we have a measure of linguistic differentiation and the corresponding date, for a language group, then such a model can be employed to predict the time depth of new language families. The calibration points used in this publication come from the paper of Holman et al. (2011). The authors use a Levenshtein distance based measure to fit lexical similarity against the time depths of the individual language families. The method of Holman et al. (2011) requires a prespecified internal classification as a necessary component to work. In other words, Holman et al.'s method is dependent on the internal structure of the language group. For instance, let us consider the calibration date for Slavic language group which is given as 1450 years old (Schenker 1995). Stated in another way, the Common Slavic language started to split up around 1450 years ago. The authors use the internal structure of the Slavic group – East, West, and South Slavic – to compute an aggregate score of the lexical divergence that occurred in this group. However, many of the language families' internal structure is not yet determined to complete satisfaction, an issue that was discussed earlier (cf. section 2.7). Hence, there is a need for a measure that can capture the divergence in a language family that reflects the family's age and is not dependent on the internal structure of the language family.

One observation that comes from studying language change is that as a language splits and develops on its own, it also develops new phonemes

different from its older stages. This shows up in the word structure of a language. For example, when sound correspondences are mapped between languages they reflect the sound change that happened in a language. For instance, the Germanic /f/ corresponding to Sanskrit /p/ reflects the sound change that occurred in Germanic from Proto-Indo-European /p/. Also, sound change can be both conditioned as well as unconditioned. A language may increase its phoneme inventory by secondary split. The surrounding environment of the phoneme contributes to the development of a new phoneme. As such, n-grams capture the environment of a phoneme.

Linguists also observed that rate of language change differs across the branches in a language family. For instance, Armenian changed when compared to Lithuanian which is supposed to have retained some of the traits from the Proto-Indo-European times. The idea behind the research presented in publication IV (chapter 10) is that counting the n-gram types found in the basic vocabulary – supposed to be resistant to lexical replacement – can be employed to gauge the amount of change that happened in a language group. Also, n-grams do not require the assumption of an internal structure of a language group that is required in Holman et al.'s analysis.

Holman et al. (2011) provide a list of 52 calibration dates for different language groups that are used to fit a linear regression model with the age of a language group as response variable and the lexical similarity (computed from lexical divergence) as predictor variable. They find that there is an inverse correlation between time depth and log of lexical similarity. I work the other way around with the size of the n-gram type inventory of a language group being the indicator of the language group's age. I employed a generalized linear model (McCullough and Nelder 1989: GLM) to fit n-gram diversity against age. The reason for not using a linear regression model comes from the nature of the data which will be explained below.

Each data point corresponds to a language group whose age is determined from archaeological, epigraphic, and historical evidence. For example, Indo-European is given as 5500 years old (without the Anatolian branch) based on the "words for wheeled vehicles" (Nichols and Warnow 2008). In reality, the model should account for the fact that 5500 years is not an exact date but an average date. For example, Western Turkic languages are supposedly 900 years old based on the evidence that the Kipchak empire spread between the 11th and 12th centuries. The calibration date is taken to be the mid point between the 11th and 12th centuries.

The Indo-European family (without the Anatolian branch) is considered to be $5000 - 6000$ years old and as such the date of 5500 is a mid-point.[23] Any statistical model should be robust to the variation in the inexactness of the dates. In fact, the magnitude of the variation in a date is an indicator of the confidence in the age of a family. In another example, the Scandinavian language group is assigned an age of 1100 years based based on the mid-point of the Viking age ($750 - 1050$ CE). It seems that historical and epigraphic dates reflect lesser uncertainty than archaeological dates.

The second issue is the non-independence of the data points. For example, the Scandinavian languages are a subgroup of the larger Indo-European family. The third issue is with regard to the amount and rates of change that happen in different families. It is generally agreed that the processes of language change that happen in different families are similar but do not have the same rate of change. For example, the Austronesian language family is spread across a large geographical area and consists of dialect chains (Blust 2011). The Austronesian languages are spoken on islands that are separated by small distances such that the languages are not geographically isolated enough to become mutually unintelligible. In contrast, the Indo-European language family is also spread across comparable geographic area but features far more mutual unintelligibility.

The ordinary linear regression works with the assumption that each data point has a constant variance σ^2. This can be thought of as each data point sampled from a normal distribution centered at the data point with a variance of σ^2. However the dates, as discussed above, show different variability. How can we capture the variability in each date? The response variable which is the age of a language family is always continuous and positive. The predictor variable (n-gram diversity) is always positive. The non-constant variability in each date can be modeled using the Gamma distribution.

The Gamma distribution has positive support and is defined using two-parameters: mean μ and shape ν. The Gamma distribution with mean μ and shape ν parameters is defined as followed:

$$\frac{1}{\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu} e^{-\frac{\nu y}{\mu}} \frac{1}{y} \tag{4}$$

The mean of the Gamma distribution is μ and its variance is μ^2/ν. The variance of each point is dependent on its mean and is a quadratic function of the mean. If a date is old, then we can expect that there is larger uncertainty as

[23] Indo-European is a language family whose age is thoroughly debated (Mallory 2013). Chang et al. (2015) cite a time period of $5500 - 6500$ years for PIE dispersal date and $5000 - 6000$ years for the Proto-Nuclear-Indo-European language (common ancestor of the non-Anatolian languages).

compared to the younger dates. The factor $1/\nu$ is known as dispersion parameter (ϕ) in GLM literature. As ϕ becomes closer to zero (or $\nu \to \infty$), the Gamma distribution becomes more symmetric (looks like a normal distribution) otherwise it is a skewed distribution. Figure 10.2 shows the scatterplot between the different n-grams and the age of language groups on a log-log scale. The lower panels give the impression that there is a linearity between n-grams and age on a log-log scale. If we try to use ordinary linear regression, the fitted dates end up being negative for younger language groups.

The GLM framework offers a way to transform the response variable without giving undesirable results. The GLM framework has the following components:

- A dispersion parameter ϕ.
- A link function that g that links the response to the predictor.

The scatterplots between age and n-grams in figure 10.2 shows that a log link function is quite appropriate. Hence, the final form of what GLM estimates, for i^{th} data point, will be:

$$g\{E(y_i)\} = b_0 + b_1 x_i \tag{5}$$

where b_0 and b_1 are coefficients of the model, x_i is the log of the number of n-grams, and y_i is the age. In our case, g is *log* function. In contrast, the ordinary linear regression estimates the following:

$$E\{g(y_i)\} = b_0 + b_1 x_i \tag{6}$$

The link between the variance and mean becomes much clearer by looking at the coefficient of variation which is defined as the ratio of standard deviation to mean. In the case of Gamma distribution it is $\sqrt{\phi}$. The overall theme in the Gamma regression is that we estimate a shape parameter for each date. Some widely used distributions such as the Chi-Square and Exponential distributions are special cases of Gamma distribution.

In publication IV (chapter 10), I tested if the predictions of the Gamma regression are influenced by the number of languages in a group, the language family itself, geographical area, and mode of subsistence. I find that the 3-grams are not biased due to the above factors. The value of ϕ is estimated to be around 0.19 which suggests that the underlying distribution for each data point is not symmetric but a skewed distribution.

The GLM procedure also outputs the standard deviation of the estimated age (also known as standard error in statistics literature). Akin to the linear

regression model, GLMs also operate with *hat* matrix (H; of order $n \times n$). A *hat matrix* is a operator that transforms a observation y_i to the predicted $\hat{y}_i$. The diagonal elements (h_{ii}) of the hat matrix have the property of *self-influence* or *leverage*, meaning that a high value of h_{ii} suggests that a data point might have a high influence over the regression. A rule of thumb is to examine those points whose leverage is greater than $2p/(n-2p)$ where p is the number of parameters (2) and n is the number of data points (52). In our case, the threshold value is 0.08. The rule of thumb is based upon the fact that the average value of h_{ii} is p/n since the sum of the all the leverages is equal to p. Hence, those leverages which are higher than twice the average leverage are considered to be potentially influential candidates.

The GLM procedure also offers what is known as *jackknife residuals*. The jackknifing procedure is equivalent to *leave-one-out* cross-validation where a model is fitted on the dataset obtained after omitting a data point; and the fitted model is employed to predict the response for the omitted observation. The jackknife residuals are defined as the difference between the predicted response value and the observed response value for the omitted data point.

The jackknifing idea is also applied to identify the actual influential data points. The identification of influential data points is achieved by estimating the parameters n times by omitting every data point exactly once. The estimated parameters for i^{th} omitted observation $(b_{0(-i)}, b_{1(-i)})$ are compared to the parameters estimated from the full dataset (b_0, b_1).[24] If the difference is quite high (close to 1) then it suggests that the points are outliers and influence the analysis.

The jackknife residuals, leverage points, and influential points are shown in figure B.3 for 3-grams. The third plot shows the high leverage points that fall on the right side of the average leverage value (represented by a dotted vertical line drawn at 0.08). The same plot also shows that Cook's statistic is not close to 1 for any of the points. The fourth plot shows Cook's statistic for each data point. The plot identifies 5 points that are farther from the rest of the points. The points are as followed: Dardic, Eastern Malayo-Polynesian, Ket-Yugh, Malayo-Polynesian, and South-west Tungusic. The Cook's statistic for each of these cases falls between $0.05 - 0.08$. The low Cook's statistic value suggests that these points do not influence the analysis.

Now, I will discuss the results of the jackknifing residuals experiment. There are two types of erring data points: those with positive deviance and the rest with negative deviance. The positive deviance refers to those points where the predicted dates is lesser than the calibration date. The negative deviance is the other way around. The highest positive deviance points are

[24] This is known as *Cook's statistic*.

Dardic, Temotu, Wakashan, and Hmong-Mein. The lowest negative deviance points are Southwest Tungusic, Romani, and Oromo. When older dates such as Benue-Congo, Indo-European, and Pama-Nyungan were tested for it is shown that the predictions for these language groups are quite close to the calibration dates. All the lowest negative deviance dates are young dates. The jackknifing procedure offers the following insights:

- Older dates can be predicted with a great accuracy. Indo-European is an example of such a prediction.
- Some data points are related through descent and the model seems not be influenced by the correlation between the data points.

I compared the prediction of the model on Austro-Asiatic – a well-researched language family – and found that it is 3700 years whereas Sidwell and Blench (2011) give a date of 4000 years. The Afro-Asiatic language family is a special case which is known to be 8000 years old and 3-gram dates are 2000 years younger than the known dates. There is a ongoing debate about the validity of large families in Africa (Blench 2013) and there seems to be a growing consensus that there are more than 4 language families in Africa. Glottolog lists 59 families (including isolates) and breaks Niger-Congo and Nilo-Saharan into smaller families.[25] Khoisan does not even figure in the classification and is also broken into subfamilies. Thus the predictions have to be revised if we take Glottolog's postulation of African families since Glottolog differs significantly from the traditional four family classification.

I also employed the 3-gram model to predict the dates for those language groups which do not have any calibration dates. The predictions for different language families is given in tables B.2–B.6. For instance, I predict the age of the Dravidian language family which is a well-established language family. The 3-gram dates give a shallow date of 2000 years which is 500 years younger than that of Krishnamurti 2003. The Austronesian language family is predicted to be around 6455 years old whereas Blust (2011: 539) gives an earliest date of 5500 years for Proto-Austronesian.

In conclusion, the 3-gram model can be employed to predict the dates for those language groups where there is not much information about the age of the language group.

[25] Accessed on 2nd May 2015.

6 SUMMARY AND FUTURE WORK

This chapter summarizes the work reported in the thesis and provides pointers to future work.

6.1 Summary

Chapter 1 places the work in part III in the context of LT and summarizes related work in CHL. Further, the chapter gives an introduction to some problems and methods in traditional historical linguistics.

Chapter 2 introduces the concepts of linguistic diversity and differences, various linguistic changes and computational modeling of the respective changes, the comparative method, tree inference and evaluation techniques, and long-distance relationships.

Chapter 3 describes various historical and typological databases released over the last few years.

Chapters 4 and 5 discusses the results in the reproduced publications from a historical linguistic viewpoint.

The publications in part III of this thesis apply some LT techniques to address some of the classical problems in historical linguistics. Most of the work reported in this thesis is carried out on the ASJP database, since the database has been created and revised with the aim of maximal coverage of the world's languages. This does not mean that the methods will not work for larger word lists such as IDS or LWT.

Chapter 7 (Wichmann, Rama and Holman 2011) examines the claim of Atkinson 2011 that human languages tend to have smaller phoneme inventories as one moves away from Africa. Based on a larger lexical database, we show that there is a small correlation (not a strong one as claimed) between language distances from Africa and the phoneme inventory sizes of the languages.

Chapter 10 (Rama 2013) provides a methodology on automatic dating of the world's languages using phonotactic diversity as a measure of language divergence. Unlike the glottochronological approaches, the explicit statistical

modeling of time splits (Evans, Ringe and Warnow 2006), and the use of Levenshtein distance for dating of the world's languages (Holman et al. 2011), the paper employs the type count of phoneme n-grams as a measure of linguistic divergence. The idea behind this approach is that the language group showing the highest phonotactic diversity is also the oldest. The paper uses generalized linear models (with the log function as link, known as Γ regression) to model the dependency of the calibration dates with the respective n-grams. This model overcomes the standard criticism of "assumption of constant rate of language change" and each language group is assumed to have a different rate of evolution over time. This paper is the first attempt to apply phonotactic diversity as a measure of linguistic divergence.

The n-gram string similarity measures applied in chapter 12 (Rama and Borin 2015) show that n-gram measures are good at internal classification whereas Levenshtein distance is good at discriminating related languages from unrelated ones. The publication also introduces a multiple-testing procedure – *False Discovery Rate* – for ranking the performance of any number of string similarity measures. The multiple-testing procedure tests whether the differential performance of the similarity measures is statistically significant or not. This procedure has already been applied to check the validity of suspected language relationships beyond the reach of the comparative method (Wichmann, Holman and List 2013).

Chapter 11 (Borin et al. 2014) shows that the edit distance based lexical comparison of word lists from South Asia groups languages into their respective language families and not into areal groups.

Chapter 8 (Rama and Kolachina 2012) correlate typological distances with basic vocabulary distances, computed from ASJP, and find that the correlation – between linguistic distances computed from two different sources – is not accidental.

Chapter 9 (Rama and Borin 2013) explore the application of n-gram measures to provide a ranking of the 100-word list by its genealogically stability. We compare our ranking with the ranking of the same list by Holman et al. (2008a). We also compare our ranking with shorter lists – with 35 and 23 items – proposed by Dolgopolsky (1986) and Starostin (1991: attributed to Yakhontov) for inferring long-distance relationships. We find that n-grams can be used as a measure of lexical stability. This study shows that information-theoretic measures can be used in CHL (Raman and Patrick 1997; Wettig 2013).

Chapter 13 (Rama 2015) on gap-weighted subsequences tests the efficiency of gap-weighted subsequences for the purpose of pair-wise cognate identification on the Indo-European database of Dyen, Kruskal and Black 1992. The paper finds that skip bigrams, when compared to skip-grams of higher lengths, perform the best at the task of pair-wise cognate identification.

6.2 Future work

The current work points towards the following directions of future work.

- Employ longer word lists such as IDS and LWT for addressing various problems in CHL.
- Apply all the available string similarity measures and experiment with their combination for the development of a better language classification system.
- Combine typological distances with lexical distances and evaluate their success at discriminating languages. Another future direction is to check the relationship between reticulation and typological distances (Donohue 2012).
- Since morphological evidence and syntactic evidence are important for language classification, the next step would be to use multilingual treebanks for the comparison of word order, part-of-speech, and syntactic subtree (or treelet) distributions (Kopotev et al. 2013; Wiersma, Nerbonne and Lauttamus 2011).
- The phonotactic diversity in publication IV (Rama 2013) can be extended to include the phylogenetic tree structure into the model. As of now, the prediction model assumes that there is no structure between the languages of a language group. A model which incorporates the tree structure into the dating model would be a next task (Pagel 1999).

Part III

Publications

7 PHONOLOGICAL DIVERSITY, WORD LENGTH, AND POPULATION SIZES

Wichmann, Søren, Taraka Rama, and Eric W. Holman. 2011. Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15: 177–197.

7.1 Introduction

Hay and Bauer (2007) report a positive correlation between population sizes and phoneme size inventories. Atkinson (2011) replicates this result, and additionally reports a negative correlation between phoneme size inventories and the distance from Africa of a given language. Neither Hay & Bauer nor Atkinson cite Nettle 1995, 1998, 1999a, who has suggested, based on a small sample of languages, that phoneme inventory sizes and mean word length are inversely related.

In this article we investigate the replicability of the results of Hay and Bauer (2007) and Nettle 1995, 1998, 1999a on a larger dataset comprising not just a few dozen languages as in the Nettle studies or a few hundred as in Hay and Bauer's study, but more than 3,000 languages carrying unique ISO 639-3 codes, a sample which represents close to one half of the world's spoken languages as defined in Lewis 2009. The languages in our sample are listed in Table S-1 of the Online Supplementary Materials for the present article. Our dataset, known as the ASJP Database and available as Wichmann et al. 2010b, consists of lists of 40 standard concepts and the corresponding words for these concepts in different languages. Different sections of the present article draw upon specific subsets of the database, as will be specified. The total sample includes languages from 109 out of the world's 121 linguistic families, 47 out of 123 isolates and unclassified languages, and 40 out of 122 creoles, mixed languages, and pidgins. Each family or non-genealogical group (such as "Unclassified" and "Creole") is represented by around one half of its members in the database.

The ASJP Database thus has the advantage of a large coverage of languages, but it is possibly disadvantaged for the purposes of the present study by containing words and not actual phoneme inventories and furthermore by a transcription procedure by which certain phonological distinctions are merged. In Section 2 we therefore describe the nature of the data further and investigate the degree to which the segments represented in the word lists (henceforth SR for "Segments Represented") are numerically proportional to phoneme size inventories. In Section 3 we correlate SR with mean word length and in Section 4 with population sizes. Patterns of residence and migration will influence the distribution of languages and therefore also the linguistic typological profiles of different world areas. It has been claimed in Atkinson 2011 that the distribution of phoneme size inventories reflects migrations pertaining to the very first movements of humans out of Africa. While it is inherently doubtful that phoneme inventory sizes can change quickly enough to "stay in tune" (i.e., correlate) with population sizes and at the same time preserve a signal many thousands of years old, there is no doubt that migrations underlie some typological distributions. In Section 5 we therefore discuss Atkinson's claim. Section 6 summarizes the findings.

7.2 SR as predictor of phonological inventory sizes

As a first step toward using the ASJP data for the study of the worldwide distribution of phoneme inventory sizes we correlate SR (segments represented in the word lists) with the segment inventory sizes of UPSID (Maddieson and Precoda 1990a), which contains information on phonological segments for 451 languages. The UPSID sample was designed to serve as a source for Maddieson 1984 and was chosen so as to be representative of each of the world's language groups, but presumably also to maximize the coverage of the variation in sound structures across the world's languages. It includes as many as 919 different phonological segments.

We would like to know how representative SRs are of given language's inventory of phonological segments. There are two obvious sources for potential discrepancies between SRs and UPSID inventories. The first is the inherent limitation of word lists: we cannot expect a short list of words to contain all of the phonological segments in a language, especially if the language has a very large segment inventory. The second source for potential discrepancies is the way in which segments are transcribed in the ASJP lists, henceforth ASJPcode. Thus, in the following paragraphs we provide some more detail on the nature of the word lists and on ASJPcode.

The word lists comprise a 40-item subset of the so-called Swadesh list, where the concepts were selected for their higher stability (Holman et al. 2008a), i.e., for the tendency for words for these concepts to only be slowly replaced by new words. As a rule of thumb the database normally only includes lists that are at least 70% complete, i.e., which contain at least 28 items on the 40-item list. Less complete lists are only included in a handful of exceptional cases where the importance of the list was judged to override the usual criterion. The concepts on the 40-item subset of the Swadesh are: BLOOD, BONE, BREAST, COME, DIE, DOG, DRINK, EAR, EYE, FIRE, FISH, FULL, HAND, HEAR, HORN, I, KNEE, LEAF, LIVER, LOUSE, MOUNTAIN, NAME, NEW, NIGHT, NOSE, ONE, PATH, PERSON, SEE, SKIN, STAR, STONE, SUN, TONGUE, TOOTH, TREE, TWO, WATER, WE, YOU (SG). It is often the case that more than one word is available for a given concept, i.e., synonyms, near-synonyms, or phonological variants. In the present study we arbitrarily use only the first item in a list of alternative forms to avoid the introduction of biases from the nature of the sources (large dictionaries vs. shorter vocabularies or the individual practices of different transcribers) and to enhance tractability of the results: using only one synonym per word allows us to equate the number of attested concepts with the number of words in a list.

One might wonder whether some other selection of concepts would be more adequate for sampling phonological segments. A larger list would obviously increase the probability that all segments of language are represented in the list, but the strength of this relationship is an empirical question.

We do find a small positive correlation ($r = .17$) between the number of words attested (which ranges from 23 to 40 with an average of 35.7) and SR for our total sample of 3,168 languages (see Table S-1 in Online Supplementary Materials). So the number of words matters, but since the frequency distribution of phonological segments presumably has a Zipfian nature in texts we would expect some sort of relation between word list size and SR which would produce diminishing returns with more words.

As for the selection of concepts one may pause to consider which sorts of words ought to be included in a list for the segments to be maximally representative of the total inventory. Since the relation between sound and meaning in language is mostly arbitrary a word for any concept can potentially contain any phoneme. But there are two special classes of words that can exhibit phonemes otherwise not attested in the standard vocabulary, namely onomatopoeic words and loanwords. The ASJP lists do not include concepts that are likely to be subject to onomatopoeia in the narrow sense of sound imitation, even if they do include concepts which in some languages are prone to sound symbolism in a broader sense, cf. Wichmann, Holman and

Brown 2010. So rare phonemes confined to onomatopoetic expression are not expected to be represented. But such phonemes are in any case often treated as marginal or as not belonging to regular segment inventories by descriptive phonologists. The lists do, however, often include loanwords, maybe on the order of 5% or so on average (currently we only have estimates available for longer versions of the Swadesh list, indicating an average of 8.5% for lists of 99 Swadesh items across a sample of 36 languages, cf. Holman et al. 2008a).

Thus, as regards the size and nature of the selection of concepts it does not seem that there is any particular reason to expect that ASJP lists could not be representative of at least a regular proportion of segment inventories. We now turn to the issue of transcription.

The transcription system, ASJPcode, was first presented in Brown et al. 2008. The system operates with 34 basic consonantal symbols and 7 vowel symbols (see tables 7.1 and 7.2). The representation of vowels is limited to at most 7 different qualities, as reflected in the 7 symbols, but in addition nasalization can be indicated by an asterisk following a vowel symbol. The 34 consonantal symbols can be combined freely to represent phonetically complex segments that are subsequently treated as single phonological units. The symbols $\sim$ and \$ follow sequences of respectively two and three consonant symbols to indicate that such sequences are to be treated as units. For instance, kw$\sim$ indicates a labialized k, and kwy\$ a labialized k with a palatal offglide. Finally, the modifier " indicates glottalization or implosion.

Because of the modifiers $\sim$ and \$, ASJPcode is quite versatile, but it also has some limitations. A relatively major limitation is the failure of the system to capture the distinction between retroflex and non-retroflex consonants, a distinction which is common in South Asia and Australia, for instance. Another limitation, especially worth noting here, is the convention according to which all click sounds are reduced to just one symbol. While this is a severe deficiency, it fortunately applies to a narrowly circumscribed set of languages only, namely those claimed to belong to the so-called Khoisan family in Lewis 2009, plus a few additional languages that have been in contact with Khoisan.

This is not the proper place to either defend or criticize ASJPcode in any major way. We would like to stress that we see no good reason for the reduction of information caused by such major limitations as the merging of some voicing distinctions or the neglect of retroflection. On the other hand, a possible different strategy of using a full equivalent of the International Phonetic Alphabet (IPA) may not be a viable alternative. It should be kept in mind that the data on which ASJP word lists are based vary in quality and suffer from overall inconsistency. The latter point is particularly important. For some languages IPA-style transcriptions are available, but the vast

ASJP symbol	Description
p	voiceless bilabial stop and fricative
b	voiced bilabial stop and fricative
m	bilabial nasal
f	voiceless labiodental fricative
v	voiced labiodental fricative
8	voiceless and voiced dental fricative
4	dental nasal
t	voiceless alveolar stop
d	voiced alveolar stop
s	voiceless alveolar fricative
z	voiced alveolar fricative
c	voiceless and voiced alveolar affricate
n	voiceless and voiced alveolar nasal
S	voiceless postalveolar fricative
Z	voiced postalveolar fricative
C	voiceless palato-alveolar affricate
j	voiced palato-alveolar affricate
T	voiceless and voiced palatal stop
5	palatal nasal
k	voiceless velar stop
g	voiced velar stop
x	voiceless and voiced velar fricative
N	velar nasal
q	voiceless uvular stop
G	voiced uvular stop
X	voiceless and voiced uvular fricative, voiceless and voiced pharyngeal fricative
7	voiceless glottal stop
h	voiceless and voiced glottal fricative
l	voiced alveolar lateral approximate
L	all other laterals
w	voiced bilabial-velar approximant
y	palatal approximant
r	voiced apico-alveolar trill and all varieties of "r-sounds"
!	all varieties of "click-sounds"

Table 7.1: ASJP consonant symbols

ASJP symbol	Description
i	high front vowel, rounded and unrounded
e	mid front vowel, rounded and unrounded
E	low front vowel, rounded and unrounded
3	high and mid central vowel, rounded and unrounded
a	low central vowel, unrounded
u	high back vowel, rounded and unrounded
o	mid and low back vowel, rounded and unrounded

Table 7.2: ASJP vowel symbols

majority of the data are available either in a practical orthography or in some regionally preferred transcription system (for instance, Africanists and Americanists tend to cater towards different sets of symbols, not all of which belong to the IPA). In this situation the use of a transcription system making fine distinctions, such as one between high and mid back rounded vowels (different *o*-sounds), would induce overdifferentiation with respect to the information usually available. For instance, given a single o-sound a source will usually just use the symbol o for the mid back rounded vowel rather than the IPA symbol for a low back rounded vowel even if the latter is phonetically more adequate. If ASJPcode were enriched with more symbols for vowel qualities, including a symbol for the low o-sound, transcribers would in the majority of cases not know which symbol to use when encountering an "o" in a source for lexical data. Thus, ASJPcode often induces a loss of information, but it usually protects against arbitrary transcription decisions. A somewhat more adequate transcription system is imaginable, but for the purpose of transcribing words of all the world's languages given the nature of the sources at hand ASJPcode cannot in a trivial manner be replaced with IPA (or ASCII equivalents such as SAMPA or the system used in UPSID).

We now estimate how strongly SR is related to UPSID segment inventory size. For this purpose we first need to match languages in UPSID with languages in ASJP. This matching cannot be perfect by any criterion except by the criterion that the source for the UPSID data should be the same as that for the ASJP data such that the two datasets could be said to derive from the same "doculect" (to use a term which has recently become current and which refers to a language variant as defined by a particular source for its description). Normally this criterion is too strict to be applicable in practice, but a looser version according to which the data should be produced by the one and the same linguist working on the same dialect can sometimes be

applied. We have applied this criterion whenever possible. Other less stringent criteria, which were used when the one just mentioned could not be applied, are (in descending order as criterial for a given decision): the data should pertain to the same geographically defined dialect; the variants should have similar names; the ASJP word list chosen from a set of otherwise equally good alternatives should be the more complete one. In a couple of cases UPSID seems to generalize over several languages or variants of languages as listed in Lewis 2009, such that an average of SRs from several ASJP lists seemed to be the most adequate point of comparison, i.e., Southern and Northern Itelmen and varieties of Dani. The result of this identification process was a matching between 392 of the 451 languages in UPSID with ASJP data. The names (in UPSID and ASJP) and ISO 639-3 codes as well as the full results of the comparisons are given in the Online Supplementary Materials (Table S-2).

Here we restrict the results of the comparisons to a statistical summery. For this purpose we exclude a single outlier, the language called !Xu in UPSID, which has a great number of click sounds that are disregarded in ASJPcode. It is the only click language in the sample.

The linear correlation between SR and UPSID inventory size, graphically displayed in figure 7.1, is a solid $r = .61$. (Here and elsewhere in the paper we use adjusted r-values.) The cone-shaped distribution of the datapoints displays a regularity in the proportion between the two variables. The average ratio of SR to UPSID segment inventory size is .817 with a standard deviation of .188. The ratios between the SR and UPSID inventory sizes across languages are uncorrelated ($r = -.05$) with the number of concepts attested in the ASJP lists. Thus, for all practical purposes we can ignore the number of attested concepts when SR is used as a proxy for total segment inventory size.

As a point of minor interest we note that SRs sometimes exceed UPSID segments in number (cf. cases where dots fall below the dotted line in figure 7.1). The main reason why the number of segments in word lists can apparently exceed the actual number of segments in a language's inventory is that transcribers and/or sources of the ASJP data may apply analyses that differ from those of UPSID, treating complex consonants as single segments through use of the transcriptional modifiers $\sim$ and \$. On the ASJP website, navigable through Wichmann et al. 2010b, the sources, the transcribed data, and even transcriber identities are available, making it possible to study such cases in more detail. For instance, for the extreme case of the language called Gbari in ASJP and Gwari in UPSID, where ASJP has 40 segments and UPSID only 26, the transcriber assumed that all combinations of a consonant symbol and the palatal glide (y in ASJPcode) are palatalized single phonemes, that the sequence *ts* is one phonological unit, and that all consonants followed

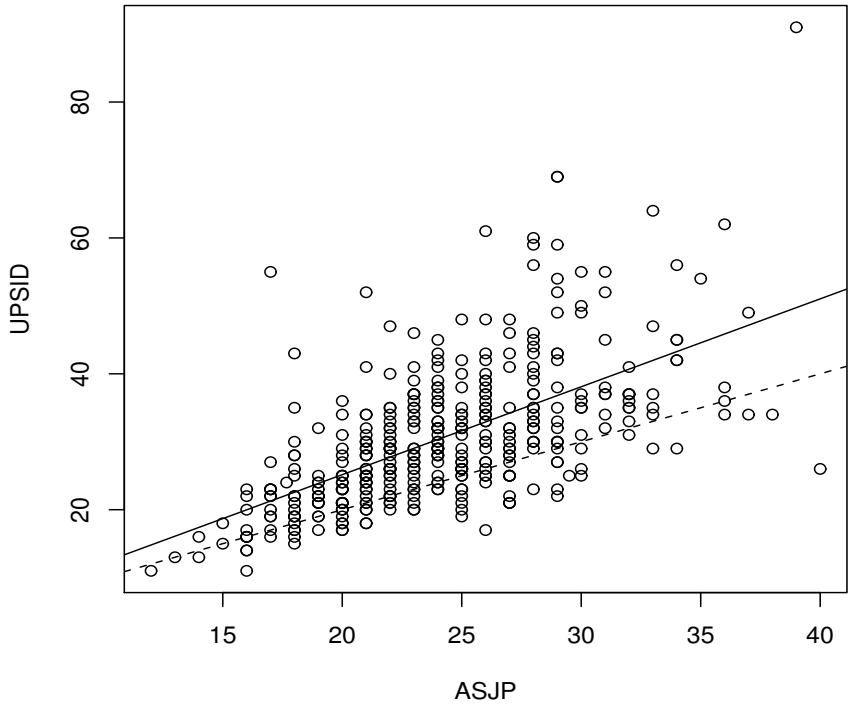


Figure 7.1: Numbers of segments in languages that are in both ASJP and UPSID. The solid line is based on a linear model predicting UPSID segment inventory size from the number of segments (SR) in ASJP word lists. The dotted line, with an intercept at (0, 0) and slope of 1, separates cases with respectively more (above the line) or fewer (below the line) UPSID segments than SRs

or preceeded by a nasal are likewise single segments. There are often possibilities for alternative analyses in phonology, and the different analyses of Gbari/Gwari could, at least in principle (we are not going to make an actual judgment in the case), both have arguments in their favor. As a matter of fact, Maddieson 1984: 6 explicitly says that when there was a choice "between a unit or sequence interpretation of, for example, affricates, prenasalized stops, long (geminate) consonants and vowels, diphtongs, labialized consonants, etc.", then there was "some prejudice in favor of treating complex phonetic events as sequences (i.e. as combinations of more elementary units)."

We conclude from the correlation in figure 7.1 that SRs in ASJP word lists are approximately proportional to segment inventory sizes to a degree where it is meaningful to use SRs as proxies for segment inventory sizes when it comes to investigating correlations with other features, such as word length, population size, and geographical distances – the topics of the next sections. If, for instance, we find a correlation between SR and average word length then this should reflect a similar or even stronger correlation between segment size inventories and average word length.

7.3 SR and word length

In two articles based on different samples of data Nettle (1995, 1998, 1999a) argues that word length is inversely correlated with the size of phonological inventories across languages. The closely related idea that the number of phonemes in a language is inversely related to the average length of morphemes is a relatively old idea in linguistics, cf. Plank 1998: 200 for references, but here we concentrate on Nettle's proposals. Nettle defines the size of a phonological inventory as the number of phonological segments available, including vowel length and tones (where the number of tones is multiplied by the number of vowels). A mean word length used for the correlation is arrived at by using 50 random dictionary entries, making sure that the dictionaries used were roughly equally sized, since larger dictionaries will tend to contain longer words on average. In Nettle 1995 ten languages from a world-wide sample are used, and in Nettle 1998 a sample of twelve languages of western Africa. We are interested in testing whether a correlation still holds up when the much larger sample of languages in the ASJP database is used. Before presenting our results we need to discuss aspects of Nettle's findings that make them different from ours in some respects even if the overall success in replicating the findings will turn out to be positive.

The are several reasons why we cannot expect the findings to be completely similar. One obvious reason, already discussed in the previous section, is that we depend on ASJPcode and a count of segment types which is often incomplete. Therefore we expect to find a weaker correlation. Another reason, which has somewhat less obvious ramifications, is that our sample is quantitatively and qualitatively different from those of Nettle. The sample of Nettle 1995 includes languages with some of the world's smallest as well as largest segment inventories, and a smattering of languages covering the range in between these extremes. The West African sample of Nettle 1998 also seems to be biased towards a coverage of the range of variation in

segment inventory sizes (this time of a particular geographical area). In contrast, our sample is not biased in any particular way, but simply contains random representatives of nearly all the world's language families where around one half of the members of each family is in the sample, as described in the beginning of Section 1 above.

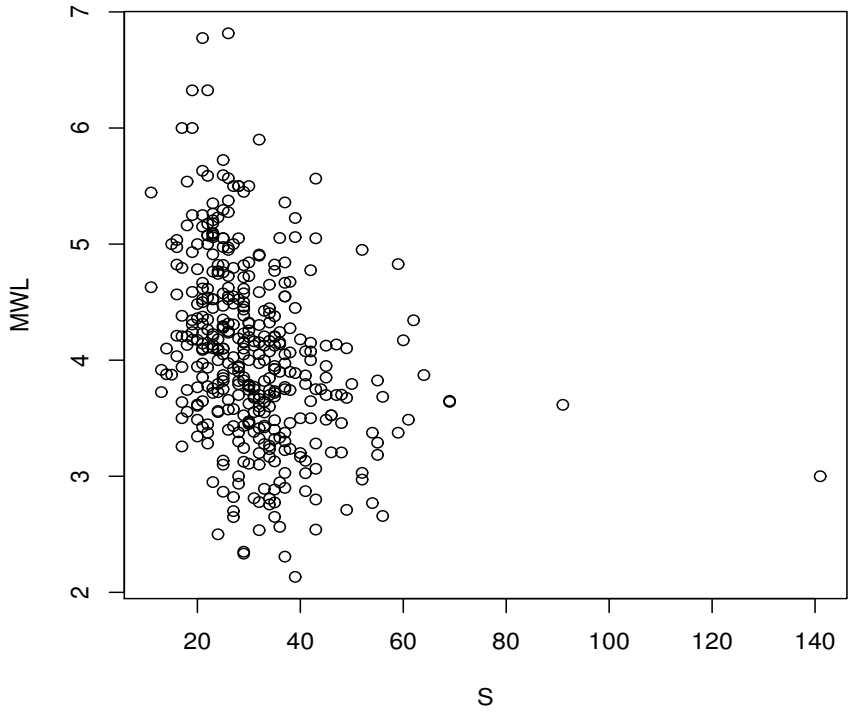


Figure 7.2: Mean word length from ASJP data plotted against segment inventory sizes from UPSID

Yet another difference is Nettle's use of segment inventory sizes (henceforth S) rather than SR. The distribution of S is positively skewed, with a long upper tail, which leads to a nonlinear relationship with mean word length (henceforth MWL) appromixating a power law. To replicate this finding with a larger set of data we can again use the UPSID data on segment inventory sizes, and the ASJP data will furnish us with information on MWL. Nettle got MWL data for each language from 50 randomly selected dictionary entries, while we will use MWL counts of the ASJP word lists. The lists used for this exercise contain from 24 to 40 words and 36.8 words on average. Thus, this way of getting

MWLs is not vastly different from Nettle's approach. The major difference is that our approach is more consistent, since we use words referring to the same concepts. In Nettle's approach there would be some added variability due to the selection procedure. For the count of S we use the number given explicitly in UPSID and do not include tonal distinctions, which is somewhat different from Nettle's count, which does include tonal distinctions. The result of plotting MWL as a function of S is shown in figure 7.2, and the data are provided in Table S-2 of the Online Supplementary Materials. As in Nettle's studies, the distribution of S is positively skewed and the relationship is nonlinear. The two outliers to the right, which account for much of the nonlinearity, are !Xu with 141 segments (in the UPSID count, which, interestingly, is different from Nettle's) and Archi with 91 segments.

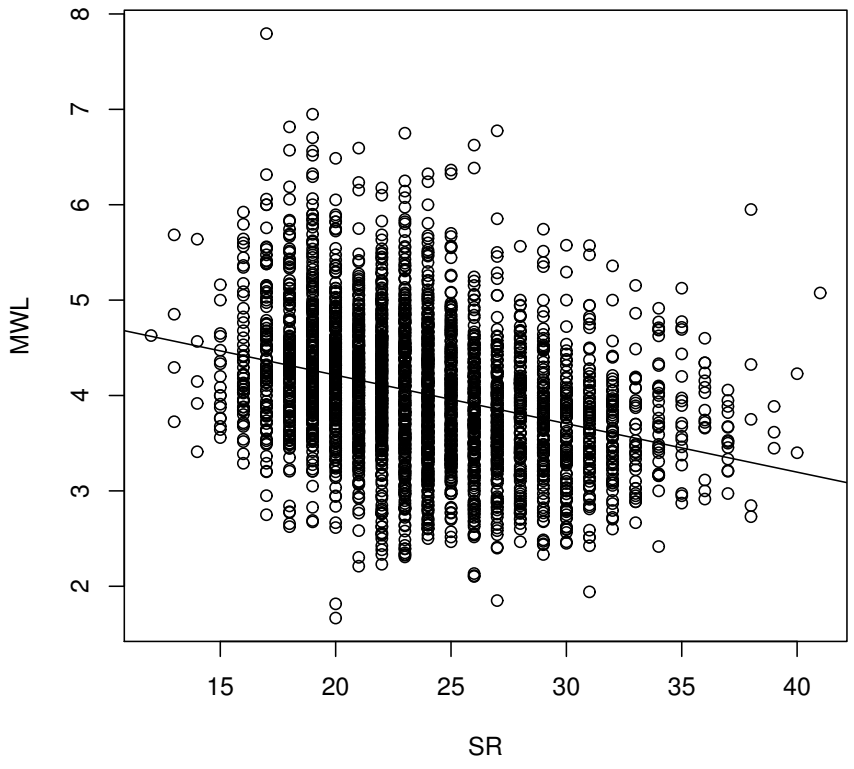


Figure 7.3: Mean word length plotted against the number of segments represented, data from ASJP

Our next step is to use all of the 3,168 ASJP languages to count MWL and SR (rather than S); see data in Table S-1 of Online Supplementary Materials. The result, displayed in figure 7.3, shows that the distribution of SR is approximately normal and the relationship is approximately linear, with $r = -.31$. Thus, for all practical purposes we can use linear regression.

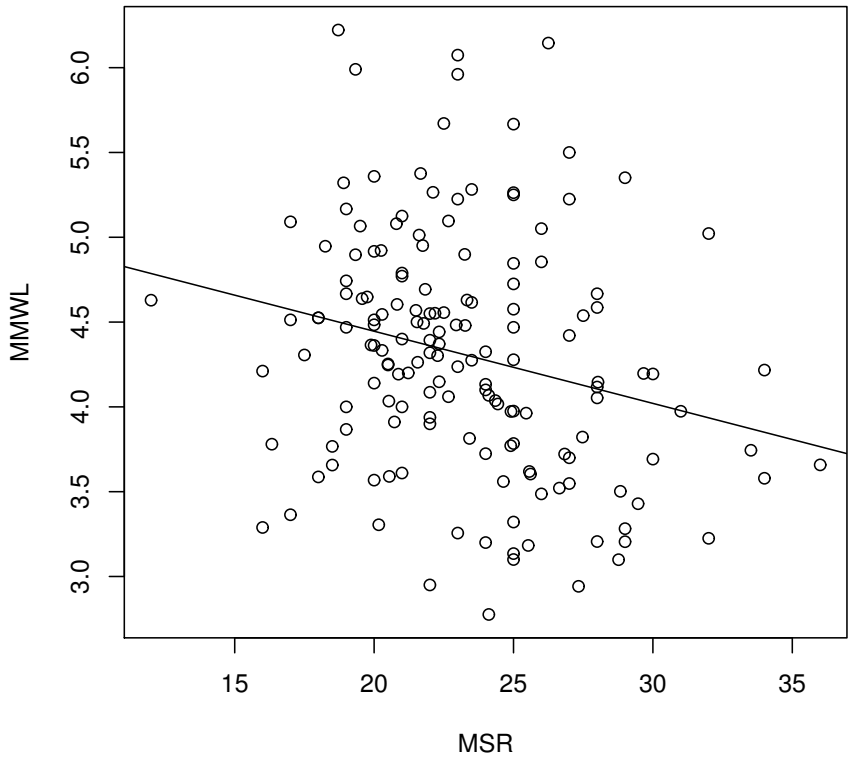


Figure 7.4: Mean word length plotted against mean SR for individual families (using ASJP data only and including all isolates and unclassified languages in the sample)

The preceding observations making reference to UPSID and Nettle 1995, 1998 had to be made in the absence of significance testing because of a sampling bias in these sources towards including the range of variation in S as well as the lack of a control for areal and genealogical effects. In contrast, a correlation found in the ASJP data can be tested for statistical significance since, as mentioned in Section 1 above, our sample is unbiased (in the sense that we are using whatever source was available) and also covers around half

of the languages in nearly all the world's linguistic families that still have members which are being spoken. To derive a conservative p-value for the correlation between SR and MWL we control for two complicating factors. The first is genealogical: languages in the same family are related through inheritance and cannot be treated as statistically independent. To deal with this problem we take averages of respectively SR and MWL within language families, and then we use families as the units of analysis in the correlations. Visual inspection of frequency histograms of SR and MWL within larger families confirmed that the distributions are approximately normal, justifying taking averages. The data for average MWL (henceforth MMWL) and SR (henceforth MSR) within families are given in Table S-3 of the Online Supplementary Materials, and are plotted in figure 7.4. The correlation is still negative: $r = -.23$ across 157 families, including isolates and unclassified languages.

The second complicating factor is geographic: processes such as language contact and migration tend to increase the similarity between geographically contiguous languages even if they are genealogically unrelated (Dryer 1989; Holman et al. 2007). In particular, there is smaller variation within geographic macro-areas as defined in Dryer 2011 than in the world at large and enough variation between macro-areas to require us to take areas into account. This point is illustrated by the boxplots shown in figures 7.5 and 7.6. (For convenience all languages of a given family are assigned to a single macro-area, even in the few cases where a family extends over two areas, such as Misumalpan and Austro-Asiatic, where the macro-area containing the majority of the languages is chosen to be representative of the family at large.)

We are now, in principle, equipped to establish the p-value for the correlation between MSR and MMWL. To control for areal effects we treat the macro-area to which each family belongs as a random effect in a linear mixed model using Baayen 2009 and Bates and Maechler 2009. The p-value is estimated using the MCMC method implemented in R as the *pvals.fnc* function of Baayen 2009. The correlation between MMWL and MSR in figure 7.4 proves to be significant ($p = .009$). This correlation may be overly conservative because it is based on all families, including those containing a single language in our sample, as well as isolates and unclassified languages. MMWL and MSR in small families are based on small samples and are therefore subject to more random sampling error than in large families, which will weaken the correlation. In fact, the correlation tends to grow when smaller families are excluded. Following the practice of Atkinson (2011) we can use the criterion that a family must contain at least two members to be included. Across the 91 families that satisfy this criterion $r = -.31$ for the

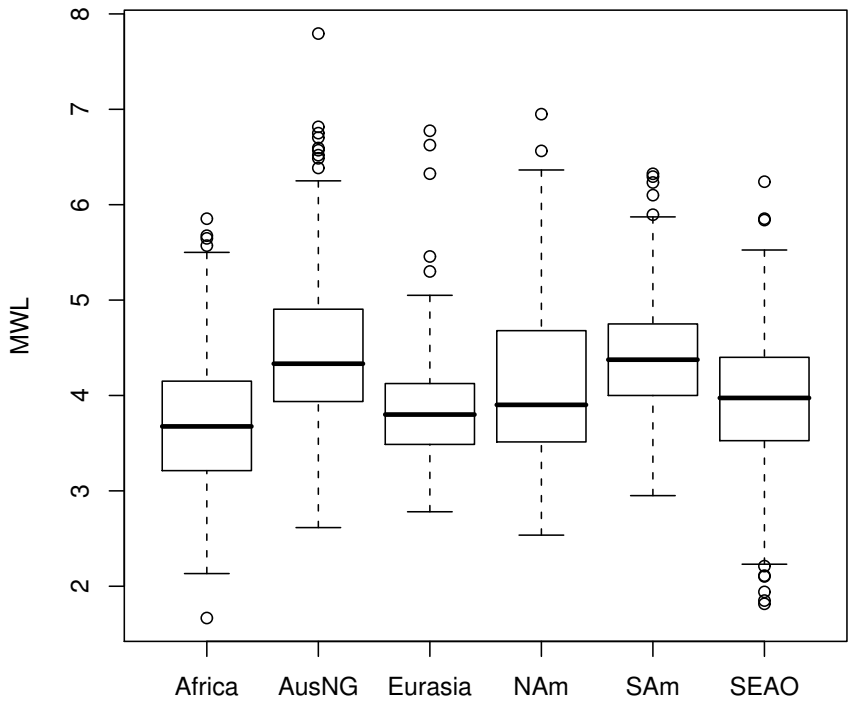


Figure 7.5: Box plots of MWL for languages pertaining to the six macro-areas of Dryer 2011

linear correlation between MMWL and MSR and $p = .0008$ for MSR as a predictor variable in the linear mixed model.

In conclusion to this section, mean word length and segment inventory sizes are significantly correlated when using a genealogically and geographically balanced sample containing close to one half of the world's languages – even when the data are somewhat impoverished because of simplified transcription and samples of segments from word lists rather than full inventories. Thus, in general, we can confirm the findings of Nettle (1995, 1998), although the size of the correlation in the larger sample is not as great as reported in his papers.

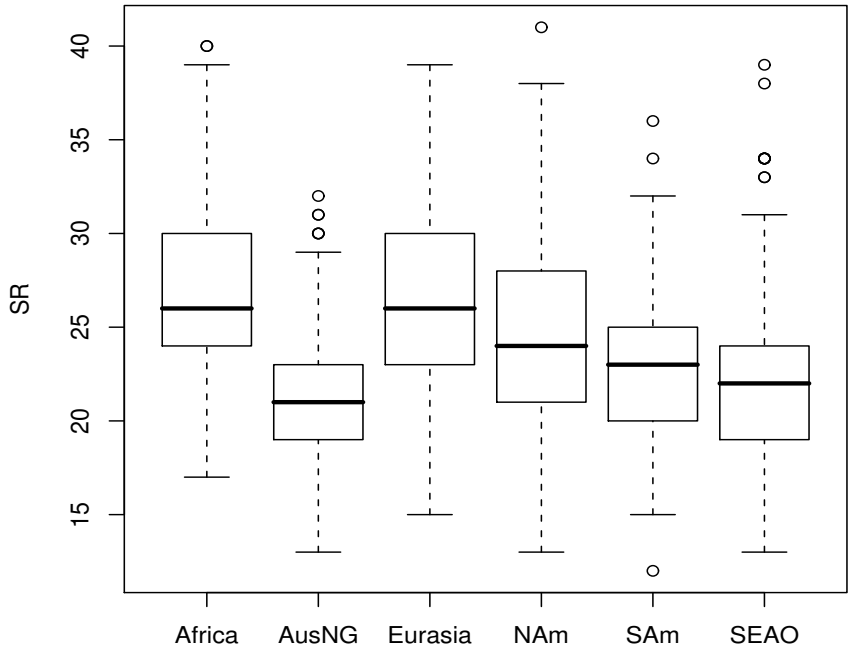


Figure 7.6: Box plots of SR for languages pertaining to the six macro-areas of Dryer 2011

7.4 SR and population sizes

Early studies of speaker population sizes and phoneme inventory sizes relied mainly on anecdotal evidence. Haudricourt (1961) discussed extra-linguistic factors that may influence inventory sizes, such as geographical isolation leading to the maintenance of large inventories, and bilingualism, which may lead to the mergers of some contrasts (cf. Labov 1994 for support), but which could also imply the introduction of new sounds from other languages (cf. also Nichols 1992). He noted that smaller populations are more likely to have a high proportion of bilinguals than larger populations, but did not explicitly state whether this should lead us overall to expect larger or smaller inventories in small populations – the effect of bilingualism can pull in both directions. Trudgill (2002) qualified this model of opposed effects further,

suggesting that it is mainly bilingualism among children which may lead to enriched inventories, whereas it is mainly adult bilingualism which causes simplification.

Hay and Bauer (2007) were the first to report a statistically validated positive correlation between speaker population sizes and phoneme inventory sizes. Their sample consists of 216 languages, and the selection ultimately derives from a textbook (Bauer 2007) where languages were chosen such as to be widely representative of different areas and linguistic families or simply to be of special interest to a linguistics student. The authors report that Spearman's $\rho = .37$ for the correlation of total phoneme inventory sizes and logarithms of population size across the total set of languages. A low value of $p < .0001$ is also given, but is not to be trusted because of failure to control for interdependence of datapoints. They also present an analysis factoring in language family as an independent variable, where families for which seven or more languages were available were treated as separate groups, and where other datapoints were lumped together in an "other" category. The result is a correlation of $r = .49$. Finally, the correlation was also tested by using means of logarithms of population sizes and phoneme inventory sizes for each family, giving $\rho = .46, p = .003$. The authors discuss possible factors such as the ones mentioned in the previous paragraph that may cause population sizes and phoneme inventory sizes to be intertwined, but refrain from pushing any particular explanation.

Atkinson (2011) finds support for Hay and Bauer 2007 and furthermore identifies an overall negative correlation between phoneme inventory sizes and the distance of languages from Africa. The two observations are brought to bear on one another by Atkinson, but presently we will focus on the first observation only, while the second will be the topic of our next section. Atkinson uses WALS (Dryer 2011) data for his study. It is to be noted that WALS operates with categorical values for segment inventory sizes rather than absolute numbers, e.g., consonant inventory sizes are put in categories from "small" to "large" with three intermediate categories. In order to arrive at values for the total inventories, Atkinson combines information from three different WALS chapters (Maddieson 2011a, b, c). For an uncontrolled correlation between phoneme size inventories and log population sizes based on 503 languages, Atkinson reports that $r = .39$ and an analysis which controls for genealogical relatedness using language family means yields $r = .47, d.f = 49, p < .001$ among families and also an effect within families. The results are similar to those of (Hay and Bauer 2007).

Following Hay and Bauer and Atkinson we replicate these results, first by finding the uncontrolled correlation for our total sample of single languages, and then by controlling for genealogical relatedness using averages over

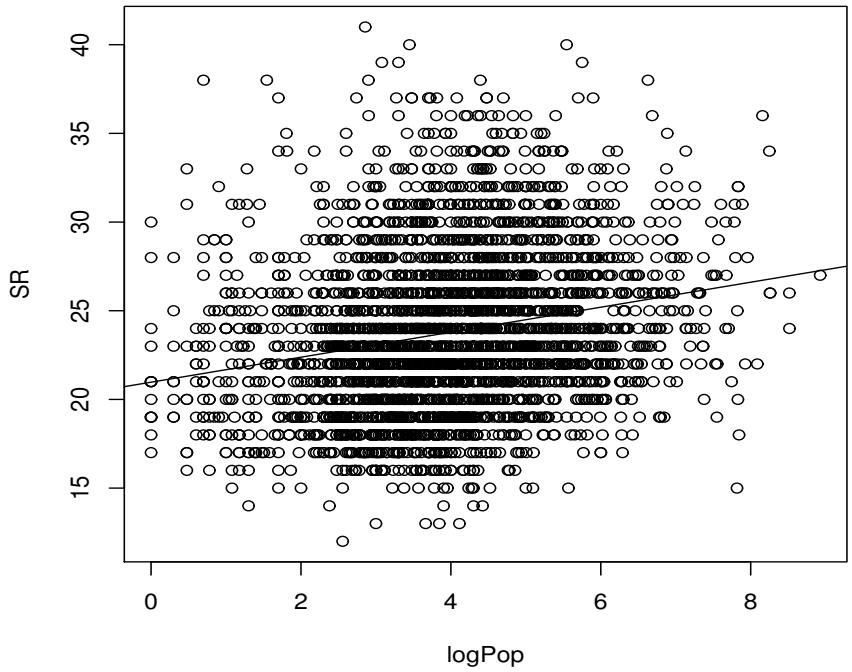


Figure 7.7: Segments represented in ASJP word lists for individual languages with one or more speakers plotted against log of population sizes from Lewis 2009

language families. The result of the first step is plotted in figure 7.7 (cf. Table S-1 of the Online Supplementary Materials for the data). Our sample of $3,153$ languages having one or more speakers is more than 14 times as large as the sample of Hay & Bauer and more than 6 times as large as that of Atkinson. We get a correlation which is weaker ($r = .236$), either because of the nature of the ASJP data or the more complete sample or a combination of these two factors.

To test the significance of the relation between and SR and the log of population sizes we now average over families. To stay on the conservative side we use all the 91 families with two or more members, as in Atkinson 2011. See Table S-3 of Online Supplementary Materials for the data. A slight gain in correlation would be gotten by using families with more than six members as in Hay and Bauer 2007, but it is not clear which criterion to apply

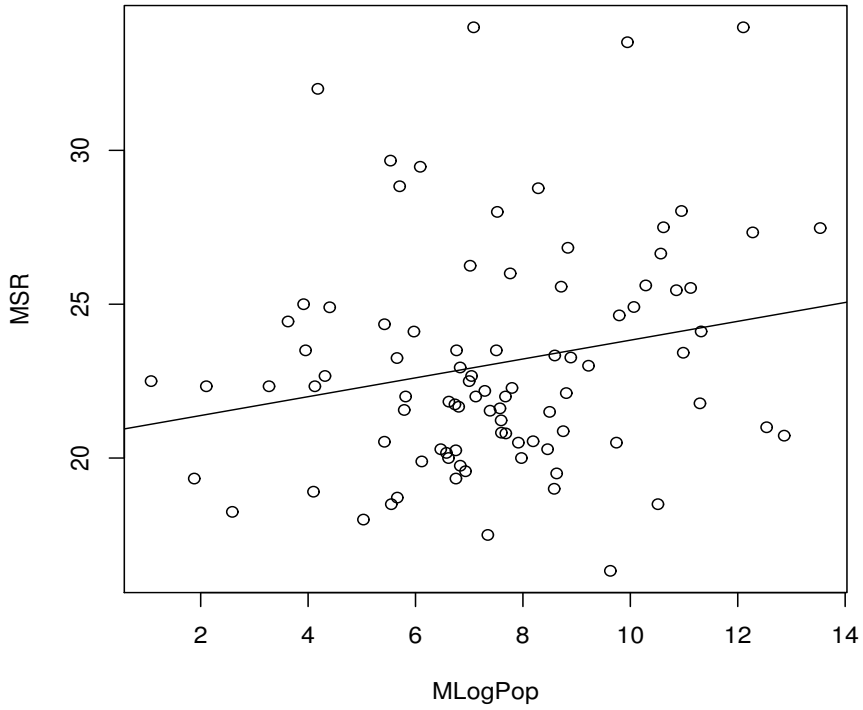


Figure 7.8: Mean of segments represented in ASJP word lists within families with 2 or more speakers plotted against the mean of log of population sizes

when excluding data points other than those representing one-member families. The result, plotted in figure 7.8, shows a correlation of $r = .18$, which is statistically significant, even if not highly so ($p = .0485$).

The fact that we get a significant correlation between population size and the number of unique segments in ASJP word lists, the latter serving as a proxy for segment inventory sizes, supports the claim of a relation between these two variables in Hay and Bauer 2007 and Atkinson 2011. Due to the nature of the ASJP data our results are somewhat inconclusive as regards the magnitude of the correlation, which is likely somewhat higher than what we find, and most likely somewhere in between our $r = .18$ and Atkinson's $r = .47$. But given the size of our sample we can firmly support the existence of the correlation first identified by Hay and Bauer (2007).

7.5 SR and geography

The purpose of this last section is to investigate the claim in Atkinson 2011 that segment inventory sizes tend to be smaller the further removed a language is from Africa. As in the preceding sections, we need to average within families as one of the requirements for establishing statistical independence of datapoints. In the following we briefly describe the details of how we proceed.

In order to choose a single geographical coordinate for each family, Atkinson used the location of a centroid language. We use a more principled approach, taking the putative homeland as inferred by the method of Wichmann, Müller and Velupillai 2010, which identifies the homeland of a given family with the language which is most diverse in the specific sense defined in their article. The difference in approaches has negligible effects since geographical ranges of language families are small in comparison to the distances between the various populated continents and Africa.

We more-or-less arbitrarily choose Addis Ababa as the point of origin of humankind within Africa. This choice does not introduce a bias in the hypothesis-testing, because the location of Addis Ababa is roughly equidistant to the coordinates that we use for three of the African families (3,367 km from Afro-Asiatic, 3,862 km from Khoisan, and 3,676 km from Niger-Congo), while having a relatively short distance (1,099 km) to the family with the smallest MSR (Nilo-Saharan, with an MSR of 24.65). A place of origin favoring the hypothesis of an inverse correlation between distance from the origin and phoneme size inventories would be closer to Khoisan and Afro-Asiatic, which have the largest MSR (respectively 28.15 and 27.36), or one could simply choose a best-fit origin as in Atkinson 2011.

Other than in these details our approach is similar to that of Atkinson. We use families for which the sample includes at least two members, and we rely on great-circle distances that are constrained to pass through the waypoints of Atkinson, i.e., Cairo, Istanbul, Phnom Penh, Bering Strait, and Panama. The major difference is in the datasets, where ours includes 91 families with a total of 3,062 languages and that of Atkinson includes 50 families with a total of 445 languages. As usual, a further difference concerns our use of SR as a proxy for phoneme inventory sizes. The results, based on the data provided in Table S-3 of the Online Supplementary Materials are plotted in figure 7.9. The correlation is a significantly negative $r = -.23$, $p = .015$.

A simple regression of the mean of the logarithm of population size and distance from Addis Ababa (henceforth "Africa") for the same language

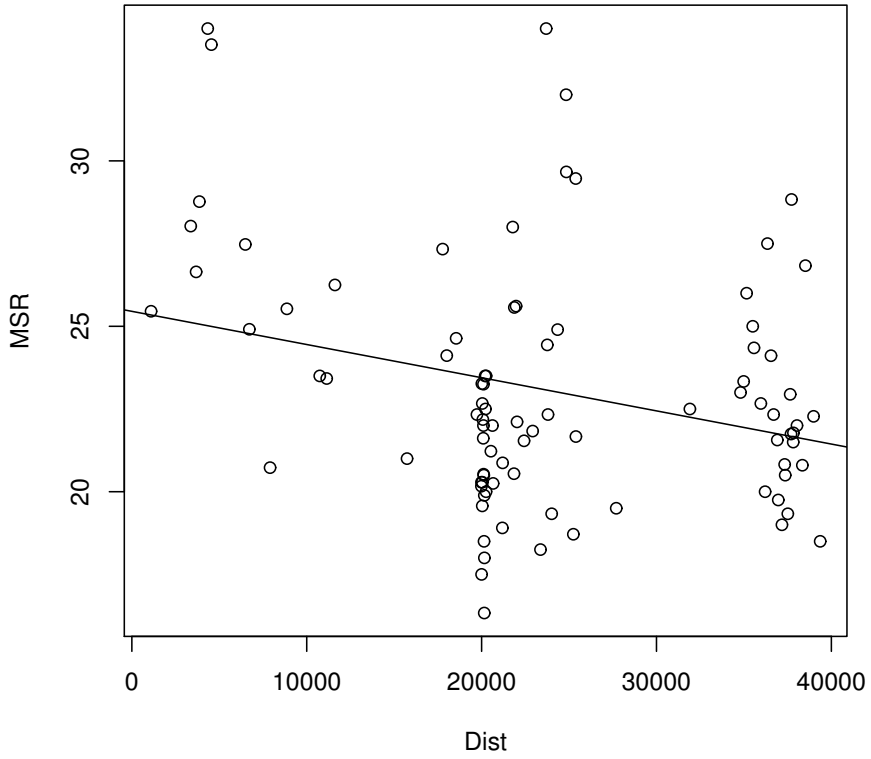


Figure 7.9: Mean of segments represented in ASJP word lists within families with 2 or more speakers plotted against the distance (in km) from Addis Ababa

families gives $r = -.34, p = .0005$.[26] This correlation is larger and more robust than that of MSR in relation to distance from Africa, which raises the question whether the inverse correlation between MSR and distance from Africa is a conspiracy of the facts that population sizes diminish with the distance from Africa (using language families as units of analysis) plus the fact that population size is correlated with MSR. The statistical significance

[26]We can confirm that population size also decreases with the distance from Africa using languages as the units of analysis. We found a correlation of $r = -.45$ based on $5,602$ languages for which geographical coordinates and population sizes were readily available to us. We do not include further details about the data on which this observation is based since we are using families as units of analysis in the present context, so the correlation involving single languages is presently largely irrelevant, even if interesting.

of the relation plotted in figure 7.9 is therefore assessed through a multiple regression with MSR as the dependent variable and distance and mean log population size as predictor variables. This produces $R^2 = .059$ and $p = .025$ for the overall multiple regression, with $p = .060$ for distance and $p = .23$ for log population. The effect of distance is not significant, suggesting that its relation to MSR may be indirect, but it is not so far from significance as to refute Atkinson's interpretation.

What we have found, then, is that an analysis similar to that of Atkinson 2011 is not inconsistent with his claim that phoneme inventory sizes really do seem to overall grow smaller as the distance from Africa increases. The question remains whether this is a secondary effect of a preference for longer words as the distance from Africa increases. This possibility is tested by another multiple regression, with MSR as the dependent variable and MMWL, distance, and mean log population size as predictor variables. This produces $R^2 = .11$ and $p = .0044$ for the overall multiple regression, with $p = .0165$ for MMWL, $p = .1413$ for distance, and $p = .5952$ for mean log population. The one significant effect confirms the negative relation between MMWL and MSR, with population as well as distance controlled. The effect of distance is again not significant, suggesting that its relation to MSR may in fact be mediated by MMWL. The role of MMWL is further investigated by a final multiple regression, with MMWL as the dependent variable and MSR, distance, and log population size as predictor variables. This produces $R^2 = .173$ and $p = .0002$ for the overall multiple regression, with $p = .0165$ again for MSR, $p = .2391$ for distance, and $p = .0180$ for log population. The new significant effect is a negative relation between population and MMWL, with MSR and distance controlled. This effect, along with the lack of a significant effect of population on MSR in the previous regression with MMWL and distance controlled, suggests that MMWL may mediate the correlation observed between population and MSR.

7.6 Discussion and conclusion

In this article we have tested different claims in the literature and were able to confirm that languages tend to have larger phoneme inventories when they have shorter words (or the other way around), that larger populations are associated with larger phoneme inventories (and therefore shorter words), and that, finally, phoneme inventories diminish with the distance from Africa. Multiple regression analyses suggest, however, that some of these relations may be indirect. How might one best account for these relations?

The relation between word length and phoneme inventory sizes is relatively straightforward. When the number of phonemes available decreases such that the probability for homonymy increases, words (i.e., lexical roots or stems) should grow longer. We can empirically observe a lower limit to the number of phonemes that a language can do with, cf. a language such as Rotokas with 11 phonemes according to Maddieson and Precoda (1990a), and most languages tend not to stay close to the limit, but rather to have a surplus of expressive means. Inversely, if for some reason – for instance through phonological erosion – a language undergoes a change towards shorter words, speakers may need to increase its inventory of phonemic distinctions. Chinese is an example where this sort of development is historically documented. Phonological erosion can also lead to increases in phoneme inventories because of clashes of segments. For instance, new diphthongs can arise through the loss of consonants that previously separated vowels or new complex consonants can arise from clusters produced through vowel epenthesis. Finally, it is also reasonable to expect that a change towards longer words (derivations, compounds) can lower the pressure on the phoneme inventory or that the acquisition of new phonemes can lower the pressure on word formation.

It is harder to explain why there is a positive correlation between population size and phoneme inventory size. Hay and Bauer (2007) refrained from insisting on an explanation, and so do we, but we would like to shift the target of the explanation. Hay and Bauer (2007) failed to cite the results of Nettle 1995, 1998, which indicate that the languages of larger populations tend to have shorter words, and our results indicate that mean word length plays a mediating role in the multiple regressions. Thus, what we need to explain is why larger populations tend to have shorter words.

As regards the inverse correlation between phoneme inventory sizes and distance from Africa, Atkinson (2011) attempts to explain this in terms of prehistorical bottlenecks. If, along migration routes, migrants pass barriers that reduce populations this could have an effect on phoneme inventory sizes, given that the two seem to be correlated. We are not quite convinced that this explanation holds. One problem with it is that prehistorical societies would have been small in any case, whether they had to pass a bottleneck or not. Moreover, a bottleneck in the normal genetic usage, which is also appropriate for the current discussion, is a historical event leading to the reduction in diversity, where only a small population is singled out for further reproduction. In a historical linguistic context this could mean the survival or passage through a migration point of a single language which could be unrepresentative of the total linguistic diversity. Such a language might have any number of phonemes. Thus, for instance, if the language(s) which made it

to the Americas happened to have large phoneme inventories then this characteristic would be transferred to modern descendants of the language(s). Indeed, the Northwest Coast of the Americas is famous for languages having large phoneme size inventories. Thus, we doubt that bottleneck effects should pull in a specific direction. Rather, they would seem to upset any kind of regularity in the development towards smaller or greater phoneme inventory sizes.

In summary, regression analysis of the ASJP data reveals a chain of three effects that are not mediated by other effects: distance from Africa is associated with smaller populations, which are associated with longer words, which are associated with fewer phonemes. The relation between distance from Africa and population can be explained by historical events such as European colonization, which was much more harmful to indigenous populations in the Americas and Australia than in Africa and Asia. As described above, processes such as avoidance of homonymy and merger of segments can account for the relation between word length and number of phonemes. The missing link is an explanation for the relation between population and word length.

8 TYPOLOGICAL DISTANCES AND LANGUAGE CLASSIFICATION

Rama, Taraka and Prasanth Kolachina 2012. How good are typological distances for determining genealogical relationships among languages? *COLING (posters)*, 975–984.

Abstract

The recent availability of typological databases such as World Atlas of Language Structures (WALS) has spurred investigations regarding its utility for classifying the world's languages, the stability of typological features in genetic linguistics and typological universals across the language families of the world. In this paper, we compare typological distances, derived from fifteen vector similarity measures, with family internal classifications and also lexical divergence. These results are only a first step towards the use of WALS database in the projection of NLP resources and bootstrapping NLP tools for typologically or genetically similar, yet resource-poor languages.

8.1 Introduction

There are more than 7000 languages in this world (Lewis 2009), which fall into more than 140 genetic families having descended from a common ancestor. The aim of traditional historical linguistics is to trace the evolutionary path, a tree of extant languages to their extinct common ancestor. Genealogical relationship is not the only characteristic which relates languages; languages can also share structurally common features such as *word order*, *similar phoneme inventory size* and *morphology*. For instance, Finnish and Telugu, which are geographically remote and yet have a agglutinative morphology. It would be a grave error to posit that two languages are genetically related due to a single common structural feature. There have been attempts in the past (Nichols 1995) to rank the stability of

structural features. Stability implies the resistance of a structural feature to change across space and time. For instance, Dravidian languages have adhered to subject-object-verb (SOV) word order for the last two thousand years (Krishnamurti 2003; Dunn, Levinson and Lindström 2008). Hence, it can be claimed that the structural feature SOV is very stable in Dravidian language family. Also, structural features have recently been used for inferring the evolutionary tree of a small group of Papuan languages of the Pacific (Dunn et al. 2005).

In the area of computational linguistics, genealogical distances between two language families have been shown to be useful for predicting the difficulty of machine translation (Birch, Osborne and Koehn 2008). However, the use of typological distances in the development of various NLP tools largely remains unexplored. Typologically similar languages provide useful leverage when working with low-resource languages. In this paper, we compare typological distances with family internal classification and also within-family lexical divergence.

The paper is structured as followed. In section 8.2, we summarize the related work. Section 8.3 lists the contributions of this work. Section 8.4 describes the typological database, lexical database and the criteria for preparing the final dataset. Section 8.5 presents the different vector similarity measures and the evaluation procedure. The results of our experiments are given in Section 8.6. We conclude the paper and discuss the future directions in Section 8.7.

8.2 Related Work

Dunn et al. (2005) were the first to apply a well-tested computational phylogenetic method (from computational biology), Maximum Parsimony (MP; Felsenstein 2004) to typological features (phonological, syntactic and morphological). The authors used MP to classify a set of unrelated languages – in Oceania – belonging to two different families. In another related work, Wichmann and Saunders (2007) apply three different phylogenetic algorithms – Neighbor Joining (Saitou and Nei 1987), MP and Bayesian inference (Ronquist and Huelsenbeck 2003) – to the typological features (from WALS) of 63 native American languages. They also ranked the typological features in terms of stability. Nichols and Warnow (2008) survey the use of typological features for language classification in computational historical linguistics. In a novel work, Bakker et al. (2009) combine typological distances with lexical similarity to boost the language classification accuracy. As a first step, they compute the pair-wise typological distances for 355 languages, obtained

through the application of length normalized Hamming distance to 85 typological features (ranked by Wichmann and Holman 2009b). They combine the typological distances with lexical divergence, derived from lexicostatistical lists, to boost language classification accuracy. Unfortunately, these works seem to have gone unnoticed in computational linguistics.

Typological feature such as phoneme inventory size (extracted from WALS database; Haspelmath et al. 2011) was used by Atkinson (2011) to claim that the phoneme inventory size shows a negative correlation as one moves away from Africa.[27] In another work, Dunn et al. (2011) make an effort towards demonstrating that there are lineage specific trends in the word order universals across the families of the world.

In computational linguistics, Daume III (2009) and Georgi, Xia and Lewis (2010) use typological features from WALS for investigating relation between phylogenetic groups and feature stability. Georgi, Xia and Lewis (2010) motivate the use of typological features for projecting linguistic resources such as treebanks and bootstrapping NLP tools from "resource-rich" to "low-resource" languages which are genetically unrelated yet, share similar syntactic features due to contact (ex., Swedish to Finnish or vice-versa). Georgi, Xia and Lewis (2010) compute pair-wise distances from typological feature vectors using cosine similarity and a shared overlap measure (ratio of number of shared features to the total number of features, between a pair of feature vectors). They apply three different clustering algorithms – k-means, partitional, agglomerative – to the WALS dataset with number of clusters as testing parameter and observe that the clustering performance measure (in terms of F-score) is not the best when the number of clusters agree with the exact number of families (121) in the whole-world dataset. They find that the simplest clustering algorithm, k-means, wins across all the three datasets. However, the authors do not correct for geographical bias in the dataset. Georgi, Xia and Lewis (2010) work with three subsets of WALS database (after applying a pruning procedure described in section 12.4). The first subset consists of 735 languages across the world. Both the second and third dataset are subsets of the first subset and consist of languages belonging to Indo-European and Sino-Tibetan language families. They divide their dataset into 10-folds and train the three clustering algorithms on 90% of the data to predict the remaining 10% of the features. Finally, the features are ranked in the decreasing order of their prediction accuracy to yield a stability ranking of the features.

[27] Assuming a monogenesis hypothesis of language similar to the monogenesis hypothesis of *homo sapiens*.

8.3 Contributions

In this article, we depart from Georgi, Xia and Lewis (2010) by not investigating the much researched topics of feature stability and the feature prediction accuracy of clustering measures. Rather, we try to answer the following questions:

- Do we really need a clustering algorithm to measure the internal classification accuracy of a language family? *Internal classification* accuracy is a measure of closeness of the typological distances to the internal structure of a language family.
- How well do the typological distances within a family correlate with the lexical distances derived from lexicostatistical lists (Swadesh 1952; Wichmann et al. 2011b), originally proposed for language classification?
- Given that there are more than dozen vector similarity measures, which vector similarity measure is the best for the above mentioned tasks?

8.4 Database

In this section, we describe WALS and *Automated Similarity Judgment Program* (ASJP), the two databases used in our experiments.

8.4.1 WALS

The WALS database[28] has 144 feature types for 2676 languages distributed across the globe. As noted by Hammarström (2009), the WALS database is sparse across many language families of the world and the dataset needs pruning before it is used for further investigations. The database is represented as matrix of languages vs. features. The pruning of the dataset has to be done in both the directions to avoid sparsity when computing the pair-wise distances between languages. Following Georgi, Xia and Lewis 2010, we remove all the languages which have less than 25 attested features. We also remove features with less than 10% attestations. This leaves the dataset with 1159 languages and 193 features. Our dataset includes only those families having more than 10 languages (following Wichmann et al. 2010a), shown in table 8.1. Georgi, Xia and Lewis (2010) work with a pruned dataset of 735 languages and two major families Indo-European and Sino-Tibetan whereas, we stick to investigating the questions in section 8.3 for the well-defined language families – Austronesian, Afro-Asiatic – given in table 8.1.

[28] Accessed on 2011-09-22.

Family	Count	Family	Count
Austronesian	150 (141)	Austro-Asiatic	22 (21)
Niger-Congo	143 (123)	Oto-Manguean	18 (14)
Sino-Tibetan	81 (68)	Arawakan	17 (17)
Australian	73 (65)	Uralic	15 (12)
Nilo-Saharan	69 (62)	Penutian	14 (11)
Afro-Asiatic	68 (57)	Nakh-Daghestanian	13 (13)
Indo-European	60 (56)	Tupian	13 (12)
Trans-New Guinea	43 (33)	Hokan	12 (12)
Uto-Aztecan	28 (26)	Dravidian	10 (9)
Altaic	27 (26)	Mayan	10 (7)

Table 8.1: Number of languages in each family. The number in parenthesis for each family gives the number of languages present in the database after mapping with ASJP database.

8.4.2 ASJP

A international consortium of scholars (calling themselves ASJP; Brown et al. 2008) started collecting Swadesh word lists (Swadesh 1952) (a short concept meaning list usually ranging from 40–200) for most of the world's languages (more than 58%), in the hope of automatizing the language classification of world's languages.[29] The ASJP lexical items are transcribed using a broad phonetic transcription called ASJP Code (Brown et al. 2008). The ASJP Code collapses distinctions in vowel length, stress, tone and reduces all click sounds to a single click symbol. This database has word lists for a language (given by its unique ISO 693-3 code as well as WALS code) and its dialects. We use the WALS code to map the languages in WALS database with that of ASJP database. Whenever a language with a WALS code has more than one word list in ASJP database, we chose to retain the first language for our experiments. An excerpt of word list for Russian is shown in table 8.2. The first line consists of name of language, WALS classification (Indo-European family and Slavic genus), followed by Ethnologue classification (informing that Russian belongs to Eastern Slavic subgroup of Indo-European family). The second line consists of the latitude, longitude, number of speakers, WALS code and ISO 693-3 code. Lexical items begin from the third line.

[29] Available at: `http://email.eva.mpg.de/~wichmann/listss14.zip`

RUSSIAN{IE.SLAVIC|Indo-European,Slavic,East@Indo-European,Slavic,EastSlavic}	
1	56.00 38.00 143553950 rus rus
1	I ya
2	you t3, v3
3	we m3
4	this iEt3
5	that to
6	who kto
7	what tato
8	not ny~E
9	all fsy~e
10	many imnogy~i

Table 8.2: 10 lexical items in Russian.

8.4.3 Binarization

Each feature in the WALS dataset is either a binary feature (presence or absence of the feature in a language) or a multi-valued feature, coded as a discrete integers over a finite range. Georgi, Xia and Lewis (2010) binarize the feature values by recording the presence or absence of a feature value in a language. This binarization greatly expands the length of the feature vector for a language but allows to represent a wide-ranged feature such as *word order* (which has 7 feature values) in terms of a sequence of 1's and 0's. The issue of binary vs. multi-valued features has been a point of debate in genetic linguistics and has been shown to not give very different results for the Indo-European classification (Atkinson and Gray 2006).

8.5 Measures

In this section, we discuss the two measures for evaluating the vector similarity measures in terms of internal classification and the computation of lexical distances for ASJP word lists. In this section, we present the 15 vector similarity measures (shown in table 8.3) followed by the evaluation measure for comparing typological distances to WALS classification. Next, we present the ASJP lexical divergence computation procedure.

Vector similarity measures

Vector similarity	
euclidean	$\sqrt[2]{\Sigma_{i=1}^{n}(v_1^i - v_2^i)^2}$
seuclidean	$\Sigma_{i=1}^{n}(v_1^i - v_2^i)^2$
nseuclidean	$\frac{\|\sigma_1 - \sigma_2\|}{2 * \|\sigma_1\| + \|\sigma_2\|}$
manhattan	$\Sigma_{i=1}^{n}\|v_1^i - v_2^i\|$
chessboard	$max((v_1^i - v_2^i) \forall i \in (1,n))$
braycurtis	$\frac{\Sigma_{i=1}^{n}\|v_1^i - v_2^i\|}{\Sigma_{i=1}^{n}\|v_1^i + v_2^i\|}$
cosine	$\frac{v_1 \cdot v_2}{\|v_1\| * \|v_2\|}$
correlation	$1 - \frac{\sigma_1 \cdot \sigma_2}{\|\sigma_1\| * \|\sigma_2\|}$

8.5.1 Internal classification accuracy

Apart from typological information for the world's languages, WALS also provides a two-level classification of a language family. In the WALS classification, the top level is the family name, the next level is genus and a language rests at the bottom. For instance, Indo-European family has 10 genera. Genus is a consensually defined unit and not a rigorously established genealogical unit (Hammarström 2009). Rather, a genus corresponds to a group of languages which are supposed to have descended from a proto-language which is about 3500 to 4000 years old. For instance, WALS lists Indic and Iranian languages as separate genera whereas, both the genera are actually descendants of Proto-Indo-Iranian which in turn descended from Proto-Indo-European – a fact well-known in historical linguistics (Campbell and Poser 2008).

The WALS classification for each language family listed in table 8.1, can be represented as a 2D-matrix with languages along both rows and columns. Each cell of such a matrix represents the WALS relationship in a language pair in the family. A cell has 0 if a language pair belong to the same genus and 1 if they belong to different genera. The pair-wise distance matrix obtained from each vector similarity measure is compared to the 2D-matrix using a special case of pearson's *r*, called point-biserial correlation (Tate 1954).

Boolean similarity	
hamming	$\#_{\neq 0}(v_1 \hat{} v_2)$
jaccard	$\frac{\#_{\neq 0}(v_1 \hat{} v_2)}{\#_{\neq 0}(v_1 \hat{} v_2) + \#_{\neq 0}(v_1 \& v_2)}$
tanimoto	$\frac{2 * \#_{\neq 0}(v_1 \hat{} v_2)}{\#_{\neq 0}(v_1 \& v_2) + \#_{=0}(v_1 \mid v_2) + 2 * \#_{\neq 0}(v_1 \hat{} v_2)}$
matching	$\frac{\#_{\neq 0}(v_1 \hat{} v_2)}{\#v_1}$
dice	$\frac{\#_{\neq 0}(v_1 \hat{} v_2)}{\#_{\neq 0}(v_1 \hat{} v_2) + 2 * \#_{\neq 0}(v_1 \& v_2)}$
sokalsneath	$\frac{2 * \#_{\neq 0}(v_1 \hat{} v_2)}{2 * \#_{\neq 0}(v_1 \hat{} v_2) + \#_{\neq 0}(v_1 \& v_2)}$
russellrao	$\frac{\#_{\neq 0}(v_1 \hat{} v_2) + \#_{=0}(v_1 \mid v_2)}{\#v_1}$
yule	$\frac{2 * \#_{\neq 0}(v_1 - v_2) * \#_{=0}(v_1 - v_2)}{\#_{\neq 0}(v_1 - v_2) * \#_{=0}(v_1 - v_2) + \#_{\neq 0}(v_1 \& v_2) * \#_{=0}(v_1 \mid v_2)}$

Table 8.3: Different vector similarity measures used in our experiments (distance computed between v_1 and v_2). In vector similarity measures, $\|\|$ represents the L_2 norm of the vector, and σ represents the difference from mean of vector (μ_1) i.e. $(v_1 - \mu_1)$. Similarly, for the boolean similarity measures, $\hat{}$ stands for the logical XOR operation between bit vectors while & and $|$ stand for logical AND and OR operations respectively. $\#_{\neq 0}(\cdot)$ stands for number of non-zero bits in a boolean vector.

8.5.2 Lexical distance

The ASJP program computes the distance between two languages as the average pair-wise length-normalized Levenshtein distance, called Levenshtein Distance Normalized (LDN) (Levenshtein 1966). LDN is further modified to account for chance resemblance such as accidental phoneme inventory similarity between a pair of languages to yield LDND (Levenshtein Distance Normalized Divided; Holman et al. 2008b). The performance of LDND distance matrices was evaluated against two expert classifications of world's languages in at least two recent works (Pompei, Loreto and Tria 2011; Wichmann et al. 2011a). Their findings confirm that the LDND matrices largely agree with the classification given by historical linguists. This result

puts us on a strong ground to use ASJP's LDND as a measure of lexical divergence within a family.

The distribution of the languages included in this study is plotted in figure 8.1.

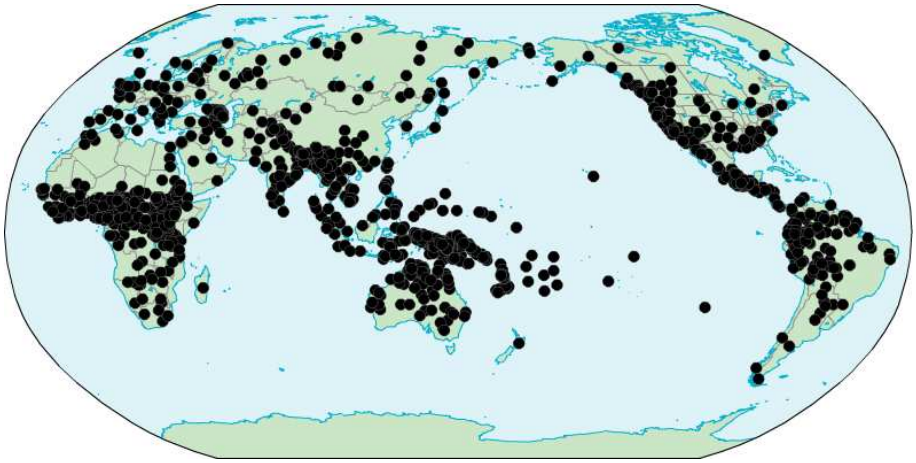


Figure 8.1: Visual representation of world's languages in the final dataset.

The correlation between typological distances and lexical distances is (within a family) computed as the Spearman's rank correlation ρ between the typological and lexical distances for all language pairs in the family. It is worth noting that Bakker et al. (2009) also compare LDND distance matrices with WALS distance matrices for 355 languages from various families using a pearson's r whereas, we compare within-family LDND matrices with WALS distance matrices derived from 15 similarity measures.

8.6 Results

In this section, we present and discuss the results of our experiments in internal classification and correlation with lexical divergence. We use heat maps to visualize the correlation matrices resulting from both experiments.

8.6.1 Internal classification

The point bi-serial correlation, r, introduced in section 8.5, lies in the range of -1 to $+1$. The value of r is blank for Arawakan and Mayan families since both families have a single genus in their respective WALS classifications. Subsequently, r is shown in white for both of these families. Chessboard

measure is blank across all language families since chessboard gives a single score of 1 between two binary vectors. Interestingly, all vector similarity measures perform well for Australian, Austro-Asiatic, Indo-European and, Sino-Tibetan language families, except for 'russellrao'. We take this result as quite encouraging, since they consist of more than 33% of the total languages in the sample given in table 8.1. Among the measures, 'matching', 'seuclidean', 'tanimoto', 'euclidean', 'hamming' and 'manhattan' perform the best across the four families. Interestingly, the widely used 'cosine' measure does not perform as well as 'hamming'. None of the vector similarity measures seem to perform well for Austronesian and Niger-Congo families which have more than 14% and 11% of the world's languages respectively. The worst performing language family is Tupian. This does not come as a surprise, since Tupian has 5 genera with one language in each and a single genus comprising the rest of family. Australian and Austro-Asiatic families shows the maximum correlation across 'seuclidean', 'tanimoto', 'euclidean', 'hamming' and 'manhattan'.

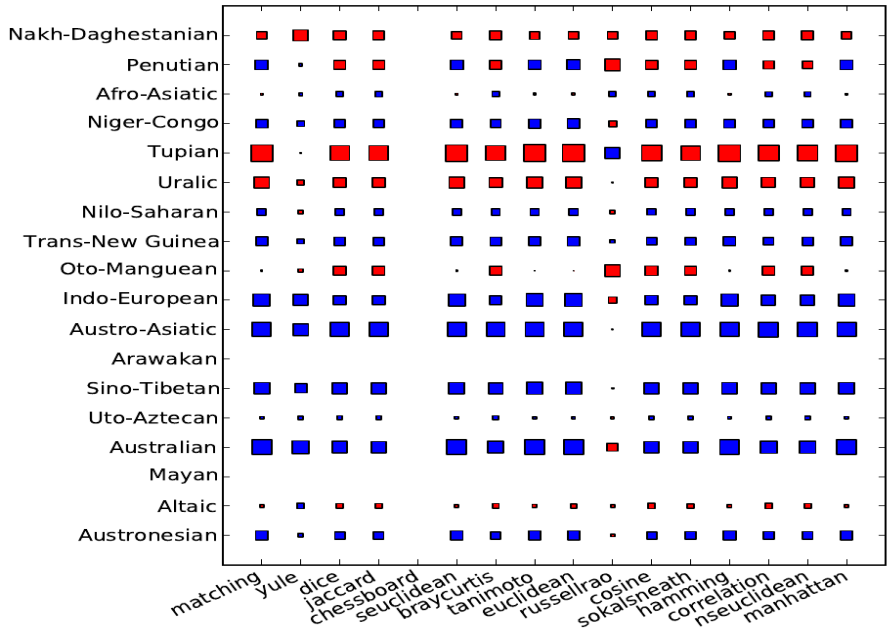


Figure 8.2: Heatmap showing the gradience of r across different language families and vector similarity measures.

8.6.2 Lexical divergence

The rank correlation between LDND and vector similarity measures is high across Australian, Sino-Tibetan, Uralic, Indo-European and Niger-Congo families. The 'Russel-Rao' measure works the best for families – Arawakan, Austro-Asiatic, Tupian, and Afro-Asiatic – which otherwise have poor correlation scores for the rest of measures. The maximum correlation is for 'yule' measure in Uralic family. Indo-European, the well-studied family, shows a correlation from 0.08 to the maximum possible correlation across all measures, except for 'Russell-Rao' and 'Bray-Curtis' distances. It is not clear why Hokan family shows the lowest amount of correlation across all the families.

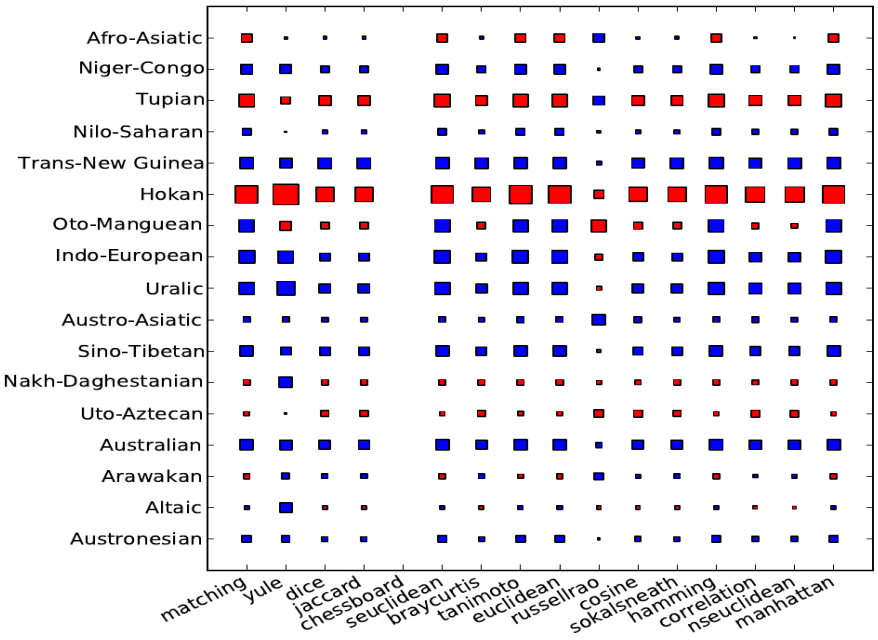


Figure 8.3: Heatmap showing the gradience of ρ across different families and vector similarity measures.

8.7 Conclusion

In summary, choosing the right vector similarity measure when calculating typological distances makes a difference in the internal classification accuracy. The choice of similarity measure does not influence the correlation

between WALS distances and LDND distances within a family. The internal classification accuracies are similar to the accuracies reported in Bakker et al. 2009. Our correlation matrix suggests that internal classification accuracies of LDND matrices (reported in Bakker et al. 2009) can be boosted through the right combination of typological distances and lexical distances. In our experiments, we did not control for feature stability and experimented on all available features. By choosing a smaller set of typological features (from the ranking of Wichmann and Holman 2009b) and right similarity measure one might achieve higher accuracies. The current rate of language extinction is unprecedented in human history. Our findings might be helpful in speeding up the language classification of many small dying families by serving as a springboard for traditional historical linguists.

9 N-GRAM APPROACHES TO THE HISTORICAL DYNAMICS OF BASIC VOCABULARY

Rama, Taraka and Lars Borin 2013. N-Gram Approaches to the Historical Dynamics of Basic Vocabulary. *Journal of Quantitative Linguistics* 21 (1): 50–64.

Abstract

In this paper, we apply an information theoretic measure, self-entropy of phoneme *n*-gram distributions, for quantifying the amount of phonological variation in words for the same concepts across languages, thereby investigating the stability of concepts in a standardized concept list – based on the 100-item Swadesh list – specifically designed for automated language classification. Our findings are consistent with those of the ASJP project (Automated Similarity Judgment Program; Holman et al. 2008b). The correlation of our ranking with that of ASJP is statistically highly significant. Our ranking also largely agrees with two other reduced concept lists proposed in the literature. Our results suggest that *n*-gram analysis works at least as well as other measures for investigating the relation of phonological similarity to geographical spread, automatic language classification, and typological similarity, while being computationally considerably cheaper than the most widespread method (normalized Levenshtein distance), very important when processing large quantities of language data.

9.1 Introduction

There are some 7,000 languages in the world (Lewis, Simons and Fennig 2013). These can be grouped into 200–400 separate language families (Lewis, Simons and Fennig 2013; Dryer 2011; Hammarström 2010). Hammarström (2010: 198) defines a language family in this way (emphasis as in the original):

- a **set of languages** (possibly a one-member set)
- with at least one **sufficiently attested** member language
- that has been **demonstrated in publication**
- to **stem from a common ancestor**
- by **orthodox comparative methodology** (Campbell and Poser 2008)
- for which there are **no** convincing published attempts to demonstrate **a wider affiliation**.

This definition implies that the set of established language families may change as research progresses, but also that there may be limits to what is knowable about the history of languages. Sometimes we find statements in the literature to the effect that a hypothesized wider affiliation is too remote to be recoverable using the traditional comparative method (Campbell and Poser 2008: ch. 9–10).

However, only a very small minority of these language families are so well-studied that there is fair consensus among experts about their internal genetic subgrouping – the structure of their family tree – at least in general outline. For the vast majority of the world's languages, the work of confirming and subgrouping established families and of combining them into more encompassing units is still very much on the wish-list of historical-comparative linguistics.

This is a vast undertaking, and to boot one pursuing a receding goal, since the world's languages are disappearing at an estimated rate of about one language every two weeks (Krauss 1992), many of them without leaving behind enough of a record so as to allow their genetic affiliations to be investigated in any detail. Here, as in other fields of scientific enquiry, we would do well to ask ourselves whether it would be possible to develop quantitative computational tools that could help experts in this endeavor. One such tool is lexicostatistics, first explicitly articulated about a half-century ago by the American linguist Morris Swadesh in a series of oft-cited papers (Swadesh 1948, 1950, 1952, 1955).

At the time – almost before computers – lexicostatistics was designed as a completely manual procedure. It relied on the manual calculation of degree of overlap – the percentage of shared cognates[30] – between short standardized lists of central and universal senses, e.g., the so-called Swadesh lists containing on the order of 200 (Swadesh 1952) or 100 items (Swadesh 1955). In lexicostatistics as originally conceived, cognacy is always determined solely by human expert judgment.

[30]In the terminology of historical linguistics, items (words, morphemes or constructions) in related languages are *cognates* if they all descend directly from the same proto-language item. This is sometimes called *vertical transmission*, as opposed to *horizontal transmission*, i.e., borrowing in a wide sense. Thus, cognacy in historical linguistics explicitly excludes loanwords.

The degree of overlap can be trivially calculated automatically once we have the information about which items are cognates and thus are to be counted as the same. Hence, the time-consuming bottleneck in lexicostatistics is the determination of cognacy, which requires considerable expertise and effort even in the case of small language families (which is not where we would expect to gain most from applying lexicostatistics in any case). Recently, some researchers have for this reason turned to approaches more amenable to automation, hoping that large-scale automatic language classification will thus become feasible.

It is important to stress at this point that such approaches are not intended as an alternative to traditional historical-comparative linguistic methodology, but rather as an addition to its toolbox. If these methods live up to expectations, they will provide an initial screening and a good first approximation of possible genetic relationships among large numbers of languages, some of which can then be singled out for more thorough investigation by human experts. The results of such investigations could then be brought back to inform and refine the automated approaches.

In most of this work, explicit cognacy judgments are replaced by an automatic calculation crucially relying on some form of (string) similarity measure, based on the assumption that, on average, cognates will tend to be more similar across languages than non-cognates. The outcome of the automated methods can be tested by comparing the automatically calculated inter-language distances to accepted language family subgroupings arrived at by the traditional comparative method, such as those provided in the *Ethnologue* (Lewis, Simons and Fennig 2013) or the *World Atlas of Language Structures* (WALS; Haspelmath et al. 2011).

In this paper, our aim is to investigate an alternative similarity measure, phoneme *n*-gram distributions, in this context, comparing it to the currently most popular measure, a variant of Levenshtein distance.

The rest of the paper is structured as follows. In the next section we give some necessary background information and a brief account of relevant related work, including the design of the ASJP database, which we use in our experiments. In the following sections, we describe and motivate the method we propose for computing item stability across language families of the world, and report on the results obtained through the application of the method, comparing them with earlier results reported in the literature. Finally, we discuss the implications of our rankings and indicate directions for further research.

9.2 Background and related work

9.2.1 Item stability and Swadesh list design

In historical linguistics, *item stability* is defined as the degree of resistance of an item to lexical replacement over time, either by another lexical item from the same language or by a borrowed lexical item. The item itself may either go out of use and disappear from the language altogether, or acquire another meaning (semantic change), i.e., move into another item slot. For example, Old English *dēor* 'animal' > *deer* (compare the Swedish cognate *djur* 'animal', which has not undergone this shift in meaning). The modern English word *animal* is a borrowing from Old French.

The quest for a *core vocabulary* (list of central lexical items; see Borin 2012 for a detailed discussion of some linguistic and computational linguistic aspects of core vocabularies) for language classification and dating of language divergence has been going on since the beginning of lexicostatistics (Swadesh 1948, 1950, 1952, 1955). The initial list of 215 items, originally presented by Swadesh in 1952, was reduced to a 100 item list in 1955. The items in the Swadesh lists supposedly represent senses universally present in human languages, and represented by words maximally resistant to lexical replacement. Unfortunately, the Swadesh lists were established mainly on the basis of Swadesh's own intuition and (considerable) professional linguistic experience, and were thus naturally limited in terms of the number of languages that could be taken into account.

Oswalt (1971) later attempted to provide more exact criteria for including a concept in the Swadesh list: (1) The cognate set[31] for the item should account for as many languages as possible. In other words, the number of cognate sets for an item should be as small as possible. (2) Cognates found in far removed languages are a stronger indicator of stability than those found in closely related languages.[32]

Together with the observation that cognates tend to be phonologically more similar than non-cognates – at least in the kind of vocabulary covered by the Swadesh lists – this opens the possibility to use (automatic) string similarity measures as proxies for (manual) cognacy judgments, thereby allowing us to test item stability on a large scale in order to investigate more

[31] The term *cognate set* refers to a set of cognate items, i.e., words in different languages going back to the same proto-language word. In working with Swadesh lists, cognates are further required to express the same sense in order for them to be in the same cognate set.

[32] Compare English *wheel* to Hindi *chakka* 'wheel', which do not reveal themselves to be cognates through visual inspection, but can nevertheless be traced back to the same Proto-Indo-European root.

objectively how well-founded Swadesh's intuitions were. To this end, Holman et al. (2008a) defined a measure – based on the phonological matches (measured using LDND; see below) between words for a single item in closely related languages (as defined in terms of WALS genera of a family; see Dryer 2011) – to rank items in a 100-item Swadesh list as to their stability and to evaluate the effect of the word-list size on automatic language classification by comparing the automatically computed inter-language distances to the genetic classification given in the WALS (Haspelmath et al. 2011) and Ethnologue (Lewis, Simons and Fennig 2013). They found that the list could be pared down to a 40-item most-stable subset without impairing the classification significantly. The resulting stability ranking of the Swadesh list items will be used in our experiment described below.

At least two recent exhaustive evaluations in automatic language classification, by Pompei, Loreto and Tria (2011) and Wichmann et al. (2011a), vindicate the use of 40-item lists across the world's language families. In both cases, a tree building algorithm (Neighbour Joining; Saitou and Nei 1987) was applied to the LDND distance matrices and the resulting trees were compared with two expert classifications (Lewis 2009; Hammarström 2010) using three different tree comparison measures, in all cases showing high agreement with the expert classifications.

9.2.2 The ASJP database

The ASJP (Automated Similarity Judgment Program) project[33] (Brown et al. 2008), comprises a group of scholars who have embarked on an ambitious program of automating the computation of similarities between languages using lexical similarity measures. The ASJP database covers a very large number of languages (more than half the world's languages in the version of the database used in the present paper). For each language present in the database, it contains a short phonetically transcribed word list based on the 100-item Swadesh list. For most of the languages this has been reduced down to the most stable 40 items, according to the empirical findings of Holman et al. 2008a, described above.

These concepts are supposed to be highly stable dischronically and therefore useful for estimating inter-language genetic distances. The ASJP program computes the distance between two languages as the average pair-wise length-normalized Levenshtein distance (Levenshtein 1966), called Levenshtein Distance Normalized (LDN). LDN is further modified to compensate for chance resemblance such as accidental phoneme inventory

[33] http://email.eva.mpg.de/∼wichmann/ASJPHomePage.htm

similarity between a pair of languages to yield LDND (Levenshtein Distance Normalized Divided; Holman et al. 2008b).

The ASJP effort began with a small dataset of 100-item lists for 245 languages. Since then, the ASJP database has been continually expanded, to include in its latest version (v. 14) more than 5,500 word lists representing well over one half of the languages of the world (Wichmann et al. 2011b). As mentioned above, most of the added word lists have aimed to cover only the 40-item most stable Swadesh subset identified by Holman et al. (2008a), and not the full 100-item list.

Each lexical item in an ASJP word list is transcribed in a broad phonetic transcription known as ASJP Code (Brown et al. 2008). The ASJP code consists of 34 consonant symbols, 7 vowels, and four modifiers, all rendered by characters available on the English version of the QWERTY keyboard. Tone, stress and vowel length are ignored in this format. The three modifiers combine symbols to form phonologically complex segments (e.g., aspirated, glottalized, or nasalized segments).

9.2.3 Earlier *n*-gram-based approaches

In quantitative approaches to historical linguistics, there are at least two earlier lines of work where character *n*-grams have been used for computing the pair-wise distances between languages, in both cases based on multilingual corpora rather than Swadesh-type word lists. Huffman (1998) compute pair-wise language distances based on character *n*-grams extracted from Bible texts in European and indigenous American languages (mostly from the Mayan language family). Singh and Surana (2007) use character *n*-grams extracted from raw comparable corpora of ten languages from the Indian subcontinent for computing the pair-wise language distances between languages belonging to two different language families (Indo-Aryan and Dravidian). Rama and Singh (2009) introduce a factored language model based on articulatory features to induce a articulatory feature level *n*-gram model from the dataset of Singh and Surana 2007. The feature *n*-grams of each language pair are compared using distributional similarity measures such as cross-entropy to yield a single point distance between a language pair.

Being based on extensive naturalistic corpus data, these studies have the considerable positive aspect of empirical well-groundedness. On the negative side, except for the study of Rama and Singh 2009, the real object of comparison – the phonology of the languages – is accessed only indirectly, through their standard orthographies, which differ in various ways, potentially distorting the cross-linguistic comparison. Another shortcoming of using

corpora for large-scale cross-linguistic investigations stems from the fact that only a small minority of the world's languages have an established written form (Borin 2009), and initiatives such as the "universal corpus of the world's languages" of Abney and Bird 2010, although of course infinitely laudable, are still a very long way from their realization.

9.3 Method

The work presented here considers a different approach from that of ASJP to investigate the individual relationship of phonological similarity with item stability. The approach in this paper is inspired by the work of Cavnar and Trenkle 1994, who use character *n*-grams for text categorization, based on their observation that the *n*-grams for a particular document category follow a Zipfian distribution. The rank of a character *n*-gram varies across documents belonging to different languages, topics and genres. Building upon this work, Dunning (1994) motivates the use of these character *n*-grams for automatic language identification and the computation of inter-language distances as well as distances between dialects.

Our motivation for conducting the present investigation has been twofold: (1) There is a general lack of comparative studies in this area, and we thus aim to contribute to the general methodological development of the field; and (2) complexity-wise, an *n*-gram-based similarity calculation is much more effective than LDND (linear vs. quadratic in the length of the input strings), and hence will scale up to much larger datasets, should the need for this arise (e.g., for comparing corpus data or full-sized dictionaries, rather than the short word lists used here).

For reasons given above, we depart from earlier *n*-gram-based approaches in that we do not use corpus data. Instead, we take advantage of the fact that the ASJP database offers an attractive alternative to corpora as the basis for massive cross-linguistic investigations. Wichmann, Rama and Holman (2011) show that the phoneme inventory sizes of 458 of the world's languages (Maddieson and Precoda 1990b) have a robust correlation ($r = 0.61$) with the number of 1-grams (supposed phonemes) extracted from the word lists for the corresponding languages in the ASJP database. Given this result, it is reasonable to assume that the phoneme *n*-grams extracted from the ASJP database give a fair picture of the phonology of the languages and consequently can be used for investigating item stability directly on the phonetic level.

All the experiments reported in this paper were performed on a subset of

version 12 of the ASJP database.[34] The database contains a total of 4,169 word lists, including not only living languages, but also extinct ones. The database also contains word lists for pidgins, creoles, mixed languages, artificial languages, and proto-languages, all of which have been excluded from the current study. Among the extinct languages, only those languages were included which have gone extinct less than three centuries ago. One might argue that phonotactic (and phonological) similarity could result from borrowing, chance or genetic affinity. We address the concern of borrowing by removing all identified borrowings from our word lists. Also, any word list containing less than 28 words (70% of the 40-item set) was not included in the final dataset. We use the family names of the WALS (Haspelmath et al. 2011) classification. Following Wichmann et al. 2010a, any family with less than ten languages is excluded from our experiments, as is any family established (only) through basic vocabulary comparison, the latter in order to avoid circularity.

Language Family$^{\text{Macro-area}}$	NOL	Language Family$^{\text{Macro-area}}$	NOL
Afro-Asiatic$^{\text{Afr}}$	9	Na-Dene$^{\text{NAm}}$	2
Algic$^{\text{NAm}}$	2	Niger-Congo$^{\text{Afr}}$	4
Altaic$^{\text{Eur}}$	2	Nilo-Saharan$^{\text{Afr}}$	2
Australian	3	Otto-Manguean$^{\text{NAm}}$	2
Austro-Asiatic$^{\text{SEAO}}$	17	Quechuan$^{\text{SAm}}$	1
Austronesian$^{\text{SEAO}}$	41	Sino-Tibetan$^{\text{Eur}}$	4
Dravidian$^{\text{Eur}}$	3	Tai-Kadai$^{\text{SEAO}}$	1
Indo-European$^{\text{Eur}}$	10	Trans-New Guinea	6
Macro-Ge$^{\text{SAm}}$	3	Tucanoan$^{\text{SAm}}$	2
Mayan$^{\text{NAm}}$	43	Tupian$^{\text{SAm}}$	3
Mixe-Zoquean$^{\text{NAm}}$	10	Uralic$^{\text{Eur}}$	3
Uto-Aztecan$^{\text{NAm}}$	3		

Table 9.1: The geographical macro-area (Haspelmath et al. 2011) of each family is indicated in superscript after the family. Afr: Africa; NAm: North America; Eur: Eurasia; SAm: South America; SEAO: South East Asia and Oceania. The rest of the language families belong to Australia-Trans New Guinea. The abbreviation for each language family is provided in brackets. NOL represents the number of languages in a family.

[34] Available on `http://email.eva.mpg.de/~wichmann/listss12.zip`.

The experiments were conducted on a dataset: corresponding to that used by Holman et al. (2008a), i.e., containing languages for which 100-item lists are available. The final dataset contains word lists for 190 languages belonging to the 30 language families listed in table 9.1.

With our proposed similarity measure, a phoneme *n*-gram profile derived from a set of similar words will contain fewer *n*-grams than one derived from a set of dissimilar words. An information-theoretic measure such as self-entropy can then be used to quantify the amount of phonological variation in a phoneme *n*-gram profile, e.g., for a Swadesh-list item across a language family. Our hypothesis is that this measure will work analogously to the LDND distance measure, and be a good, computationally cheaper substitute for it.

The phoneme *n*-gram profile for a language family is computed in the following manner. A phoneme *n*-gram is defined as the consecutive phoneme segments in a window of a fixed length *n*. The value of *n* ranges from one to five. All the phoneme 1-grams to 5-grams are extracted for a lexical item in an item list. All *n*-grams for an item, extracted from word-lists belonging to a family, are merged, counted and sorted in order of descending frequency. This list constitutes the *n*-gram profile for the item. In the next step, the relative frequency of each *n*-gram in an *n*-gram profile for an item is computed by normalizing the frequency of a phoneme *n*-gram by the sum of the frequencies of all the *n*-grams in an item's *n*-gram profile. This corresponds roughly to the length normalization step in the calculation of LDND. It can be summarized as in (7), where f^i_{ngram} denotes the frequency of the i^{th} *n*-gram and *S* denotes the size of the *n*-gram profile for an item.

$$rf^i_{ngram} = \frac{f^i_{ngram}}{\sum_{i=1}^{S} f^i_{ngram}} \tag{7}$$

Given this background, the self-entropy of the k^{th} item's *n*-gram profile can defined as in (8):

$$H^k_{item} = -\sum_{i=1}^{S} rf^i_{ngram} \cdot log(rf^i_{ngram}) \tag{8}$$

The self-entropy $H(\cdot)$ is further scaled by raising it to the power of *e* to provide better resolution. Since self-entropy $H(\cdot)$ measures the amount of divergence in the phoneme *n*-gram profile for an item, the items can be ranked relatively in terms of the ascending order of self-entropy averaged (weighted by the size of the family[35]) across the families.

[35]We use weighted average to factor out the effect of sample size of each language family. We tried averaging using the number of families and total number of languages present in the

9.4 Results and discussion

Table 9.2 shows the 100 items ranked in decreasing order of stability, as indicated by the phoneme *n*-grams method. A Spearman's rank correlation ρ between the ranks given in table 9.2 and the ranks given by Holman et al. (2008a) (listed in column *H08* of table 9.2) is 0.63 ($p < 0.001$). The correlation is quite robust and highly significant. This correlation suggests that the *n*-gram-based ranking of the 100-item list is highly similar to the ASJP ranking based on LDND.

The ASJP 40-item list (actually 43, since the Swadesh list senses 'rain', 'bark' and 'kill' are instantiated with the same lexical items as 'water', 'skin' and 'die' in many languages; hence, the reduced ASJP list covers ranks down to 43) has 35 items in common with the *n*-gram method. One simple way to test if the intersection is by chance is to run a 1000-trial simulation by selecting two random samples of 43 items from a 100-item list and counting the number of times that both the samples have items in common. Such a test showed that the result is significant.

This agreement of our rankings with that of ASJP puts us on a strong footing for the enterprise of automated language classification, as it implies that a similarity measure based on phoneme *n*-grams is a good alternative to LDND.

There are at least two shorter lists – of length 35 and 23 – proposed by Starostin (1991), attributed to Yakhontov, and Dolgopolsky (1986), both specially designed for identifying relationships between remote languages and looking past the time-depth ceiling imposed by the traditional comparative method (Kessler 2008; Campbell and Mixco 2007), and consequently aspiring to identify maximally stable items across languages.[36] The 100-item Swadesh list lacks three items, 'nail', 'tear/drop' and 'salt', present in Dolgopolsky's 23-item list. Our 40-item list has 17 items in common with the 23-item list. Yakhontov's 35-word list contains the items 'salt', 'wind', 'year' which are not present in Swadesh's 100-item list, but are in the 200-item list. Our 40-item list has 24 items in common with Yakhontov's list. We conclude our comparison with the shorter word-lists by noting that our method places 'this', 'who', 'what' and 'give' among the top 43, present in Yakhontov's list whereas, ASJP places them after 43. The items 'not' and 'who' appear in the 23-item list of Dolgopolsky but do not appear in the ASJP 40-item list.

100-word list sample. Both the averaging techniques correlate highly ($\rho > 0.92$, $p < 0.001$) with the weighted average.

[36] Dolgopolsky (1986) arrived at the 23-item list by comparing 140 languages belonging to ten families.

Rank	H08	#/Item	Stability	Rank	H08	#/Item	Stability
1	6	*1/I^{D,Y}	101.953	51	20	*44/tongue^{D,Y}	325.452
2	1	*22/louse^{D,Y}	111.647	52	96	49/belly	331.196
3	11	*23/tree	153.598	53	41	*96/new^{Y}	346.665
4	8	*40/eye^{D,Y}	157.787	54	89	65/walk	348.355
5	16	*51/breasts	166.545	55	70	37/hair	349.821
6	38	*54/drink	169.873	56	54	79/earth	351.177
7	5	*61/die^{D,Y}	175.367	57	86	35/tail^{Y}	358.716
8	43	*72/sun^{D,Y}	177.999	58	32	*95/full^{D,Y}	359.211
9	47	70/give^{Y}	178.558	59	28	*18/person	372.306
10	27	*34/horn^{D,Y}	184.285	60	64	83/ash	373.392
11	2	*12/two^{D,Y}	204.345	61	53	38/head	375.332
12	73	4/this^{Y}	211.777	62	80	17/man	380.891
13	40	27/bark	216.259	63	85	84/burn	384.256
14	33	*66/come	217.405	64	15	*43/tooth^{D,Y}	385.485
15	3	*75/water^{D,Y}	219.858	65	82	29/flesh	387.652
16	13	*100/name^{D,Y}	222.516	66	91	10/many	390.966
17	66	55/eat	223.821	67	79	97/good	398.742
18	30	*11/one^{Y}	225.599	68	83	50/neck	399.049
19	12	*19/fish^{Y}	227.537	69	98	93/hot	400.909
20	17	*2/you^{D,Y}	230. 4	70	45	32/grease	406.988
21	68	6/who^{D,Y}	236.126	71	95	63/swim	408.921
22	9	*48/hand^{D,Y}	236.955	72	63	56/bite	411.407
23	35	*86/mountain	250.578	73	84	71/say	412.343
24	4	*39/ear^{D,Y}	259.692	74	67	33/egg^{Y}	415.863
25	42	*21/dog^{Y}	263.383	75	75	16/woman	418.379
26	24	76/rain	264.581	76	10	*58/hear	421.242
27	36	*82/fire^{Y}	265.697	77	59	60/sleep	438.326
28	14	*77/stone^{Y}	268.448	78	44	64/fly	440.443
29	26	*30/blood^{D,Y}	268.841	79	25	62/kill	447. 78
30	37	*3/we	270.771	80	78	69/stand	456. 98
31	21	*28/skin	271.754	81	50	90/white	461.035
32	31	*41/nose^{Y}	275.397	82	22	*92/night^{D}	464.536
33	18	*85/path	281.417	83	97	13/big	467.197
34	88	5/that	281.823	84	61	26/root	485.709
35	29	*47/knee	283.865	85	65	87/red	490.751
36	7	*53/liver	289.086	86	94	80/cloud	493.499
37	74	24/seed	291. 34	87	51	89/yellow	496.966
38	19	*31/bone^{Y}	291.445	88	69	99/dry	499.142

39	39	*57/see	293. 99	89	77	14/long	500.151
40	55	46/foot	295.059	90	58	88/green	522.136
41	60	7/what^{Y}	295.904	91	76	98/round	525.012
42	72	8/not^{D}	297.628	92	87	78/sand	527.829
43	23	*25/leaf	297.915	93	93	59/know^{Y}	527.866
44	46	73/moon^{Y}	308.394	94	34	*74/star	558.447
45	52	20/bird	314.281	95	100	15/small	597.591
46	49	36/feather	315.486	96	81	94/cold	598.111
47	57	42/mouth	318.221	97	56	91/black	602.131
48	71	81/smoke	318.681	98	99	67/lie	619.404
49	48	52/heart^{D}	320.086	99	90	68/sit	620.247
50	62	45/claw	323.965	100	92	9/all	679.997

Table 9.2: The items are presented in the ranking given by the *n*-grams (Rank). The second column (H08) provides the corresponding ranking of Holman et al. (2008a). The Swadesh list number/item is found in the third column, where the * symbol denotes an item present in the reduced 40-item ASJP list. Superscripts D and Y indicate membership in the lists of Dolgopolsky (1986) and Starostin (1991: attributed by Starostin to Yakhontov), respectively.

9.5 Conclusions

In summary, the item stability ranks derived from *n*-gram analysis largely agree with the item stability ranks based on phonological matches found by Holman et al. (2008a) using LDND as the similarity measure. This result suggests that phoneme *n*-grams work at least as well as other string similarity measures – e.g., LDND – for investigating the relation of phonological similarity to geographical spread, automatic language classification, and typological similarity. At the same time, *n*-gram analysis is cheaper in terms of computational resources – the fundamental comparison step has linear complexity, against quadratic complexity for LDND – which is important when processing large quantities of language data.[37]

A topic in need of future research is a continuation of the methodological strand of the work presented here, in the form of a more encompassing comparison of different string similarity measures for automated lexicostatistics. There is also the general issue of whether the "classical" Swadesh lists are the optimal point of departure for identifying the most stable lexical items across the languages of the world, as has been (tacitly or

[37] The LDND program takes about one hour to compute the inter-language distances whereas, the *n*-gram analysis takes less than two minutes.

explicitly) assumed in most previous work (with Dolgopolsky 1986 forming a notable exception in this regard; see also Borin 2012 for a more in-depth discussion of this issue), or whether even more stable items could be found by looking at the matter with fresh eyes, perhaps using text corpora.

10 PHONOTACTIC DIVERSITY AND TIME DEPTH

Rama, Taraka 2013. Phonotactic Diversity Predicts the Time Depth of the World's Language Families. *PloS one 8.5:e63238*.

Abstract

The ASJP (Automated Similarity Judgment Program) described an automated, lexical similarity-based method for dating the world's language groups using 52 archaeological, epigraphic and historical calibration date points. The present paper describes a new automated dating method, based on phonotactic diversity. Unlike ASJP, our method does not require any information on the internal classification of a language group. Also, the method can use all the available word lists for a language and its dialects eschewing the debate of 'language' and 'dialect'. We further combine these dates and provide a new baseline which, to our knowledge, is the best one. We make a systematic comparison of our method, ASJP's dating procedure, and combined dates. We predict time depths for world's language families and sub-families using this new baseline. Finally, we explain our results in the model of language change given by Nettle.

10.1 Introduction

Glottochronology, as introduced by Swadesh (1952, 1955), is a method for estimating the split/divergence time of two phylogenetically related languages from their common ancestor. It makes use of Swadesh lists, which are short lists, usually 100—215 items of core vocabulary, supposed to be resistant to borrowing and is universal and culture-free.

Core vocabulary is supposedly more resistant to lexical replacement than other vocabulary items. There is an assumption of a universal constant rate of lexical change over time. The time depth of the point of split between two

languages is proportional to the logarithm of lexical similarity. The lexical similarity between two languages is measured as the percentage of cognates, C, shared between the pair of languages. The time depth is estimated in units of 1000 years using the following formula.

$$t = \frac{log\ C}{2\ log\ r} \tag{9}$$

The constant r is experimentally determined by Lees (1953) using 13 control cases.

Glottochronology was heavily criticized for several reasons, especially the following ones:

- The composition of the core vocabulary list is not objective. Only recently, in Holman et al. 2008a; Petroni and Serva 2011 was the assumption of stability of the core vocabulary tested quantitatively for the worldwide language families.
- The rate of lexical replacement is not constant across different families or within the families. As demonstrated in Bergsland and Vogt 1962, Icelandic has a relatively lower rate of lexical change and East Greenlandic Eskimo has a higher rate of lexical change than assumed by Lees (1953).

The related work in the field of computational historical linguistics is described in the next subsection.

10.1.1 Related work

The last decade has seen a surge in the number of papers published in historical linguistics applying computational and statistical methods. This literature can be broadly classified into two areas.

One area of work, represented by Wichmann et al. (2010a), Holman et al. (2008a), Bakker et al. (2009), Holman et al. (2011), Ringe, Warnow and Taylor (2002), and Gray and Atkinson (2003) focuses on collecting word lists for various language families for attacking classical historical linguistics problems such as dating, internal language classification, and lexical stability.

The other area of work, represented by papers such as Wichmann and Holman (2009a), Wichmann (2010b), Nettle (1999a), Wichmann, Müller and Velupillai (2010), Hammarström (2010), and Atkinson (2011) is characterized by the application of quantitative methods to seek answers to questions also involving socio-historical processes, including the relations

between language diversity, human population sizes, agricultural patterns and geographical origins of languages. It should be noted that this classification is not strictly mutually exclusive (see Wichmann 2008 for a survey of the computational, statistical and inter-disciplinary work on language dynamics and change). Of the several works cited above, those of Wichmann et al. 2010a, Serva and Petroni 2008, Holman et al. 2011 are relevant to this paper.

Gray and Atkinson (2003) date the Indo-European family as 8000 years old using a penalized minimum likelihood model which supports the Anatolian hypothesis of language spread. They use a binarily encoded character matrix (presence/absence of a cognate for a language; judged by comparative method) for Indo-European from Dyen, Kruskal and Black 1992 for inferring the phylogenetic tree and dating its nodes.

A completely different approach is taken by the ASJP consortium for the automated dating of the world's language families. ASJP[38] is a group of scholars who have embarked on an ambitious program of achieving an automated classification of world's languages based on lexical similarity. As a means towards this end the group has embarked upon collecting Swadesh lists for all of the world's languages. The database is described in the subsection ASJP Database below.

Holman et al. (2011) collected calibration points for 52 language groups from archaeological, historical and epigraphic sources. The intra-language group lexical similarity was computed using a version of the Levenshtein distance (LD). Levenshtein distance is defined as the minimum number of substitution, deletion and insertion operations required to convert a word to another word. This number is normalized by the maximum of the length of the two words to yield LDN, and finally the distance measure used, LDND (LDN Double normalized), is obtained by dividing the average LDN for all the word pairs involving the same meaning by the average LDN for all the word pairs involving different meanings. The second normalization is done to compensate for chance lexical similarity due to similar phoneme inventories between unrelated languages.

Now, we describe the computation of average lexical similarity for a intra-language group using the Scandinavian calibration point. The Scandinavian language group has two sub-groups: East Scandinavian with 5 word lists and West Scandinavian with 2 word lists. The internal classification information is obtained from *Ethnologue* (Lewis 2009). The ASJP procedure sums the LDND of the 10 language pairs and divides them by 10 to yield an average LDND for Scandinavian language group. Then, they fit a ordinary least-squares regression model with average lexical similarity as a predictor

[38] `http://email.eva.mpg.de/$~$wichmann/ASJPHomePage.htm`

and time depth as the response variable. The regression yields a highly robust correlation of $-.84$. Finally, they use the fitted regression model to predict a language group's ancestral time depth for different language families across the world.

Serva and Petroni (2008) were the first to use LD to estimate the time-depth of a language family. But their experiments were focused on dating the root of the Indo-European tree. They primarily use IE database (Dyen, Kruskal and Black 1992) – augmented by some of their own data – for their experiments.

10.2 Materials and Methods

10.2.1 ASJP Database

The ASJP database (Wichmann et al. 2010b; Expanded versions of the ASJP database are continuously being made available at [39]) has 4817 word lists from around half of the languages of the world including creoles, dialects, artificial languages and extinct languages. We work with the version 13 database for comparability with the results given by the ASJP dating procedure. A language and its dialects is identified through a unique ISO 639-3 code given in *Ethnologue* (Lewis, Simons and Fennig 2013). The database also contains the languages' genetic classification as given in WALS (Haspelmath et al. 2011) and *Ethnologue* (Lewis, Simons and Fennig 2013). The database has a shorter version – the 40 most stable meanings empirically determined by Holman et al. (2008a) – of the original Swadesh list. A word list for a language is normally not entered into the database if it has less than 70% of the 40 items. For our experiments, we use a subset of the data obtained by removing all the languages extinct before 1700 CE.

The word lists in ASJP database are transcribed in ASJPcode (Brown et al. 2008). ASJPcode consists of characters found on a QWERTY keyboard. ASJPcode has 34 consonant symbols and 7 vowel symbols. The different symbols combine to form complex phonological segments. Vowel nasalization and glottalization are indicated by $*$ and ", respectively. Modifiers $\sim$ and \$ indicate that the preceding two or three segments are to be treated as a single symbol.

[39] `http://email.eva.mpg.de/~wichmann/EarlierWorldTree.htm`

10.2.2 ASJP calibration procedure

The motivation for and the details of the ASJP calibration procedure is outlined in this section. There are at least three processes by which the lexical similarity between genetically related languages decreases over time. Shared inherited words (cognates) undergo regular sound changes to yield phonologically less similar words over time (e.g. English/Armenian *two* ∼ *erku* 'two'; English/Hindi *wheel* ∼ *chakra* 'wheel'). Words can also undergo semantic shift or are replaced through copying from other languages causing a decrement in the lexical similarity between related languages. LDND is designed specifically to capture the net lexical similarity between languages related through descent.

The ASJP's date calibration formula is similar to that of glottochronology (9). Eqn. 9 implies that the ancestral language is lexically homogeneous at $t = 0$. This formula is modified to accommodate lexical heterogeneity of the ancestral language at time zero by introducing s_0, representing average lexical similarity at $t = 0$ of the language groups' ancestral language. The cognate proportion C is replaced by the ASJP lexical similarity defined as 1−LDND. The formula then looks as in (2):

$$t = (log\ s - log\ s_0)/2\ log\ r \tag{10}$$

The values of s_0 and r are empirically determined by fitting a linear regression model between the 52 language groups' time depth (t) and their lexical similarity (s). The intra-language group similarity is defined as the average pairwise lexical similarity between the languages belonging to the coordinate subgroups at the highest level of classification. Eqn. 10 and the negative correlation implies that log lexical similarity has an inverse linear relationship with time depth.

The next subsection describes our findings on the relation between language group diversity and the age of the group.

10.2.3 Language group size and dates

As mentioned earlier, the ASJP consortium (Holman et al. 2011) collected common ancestor divergence dates for 52 language groups, based on archaeological, historical or epigraphic evidence. Written records can be used to determine the date of divergence of the common ancestral language reliably. The recorded history of the speakers of the languages can be used to determine the divergence dates based on major historical events. Since written

records do not exist for temporally deep language families, the date for the common ancestor must often be inferred from archaeological sources.

Archaeological dates can be determined on the basis of traceability of the proto-language's reconstructed words to excavated material objects. Dates can also be inferred if loanwords can be correlated with historical or archaeological events. The process of compiling calibration points was extremely careful and archaeological calibration points were only included if they were non-controversial. Specifically, any glottochronologically determined date was excluded from the sample.

A description of the sources of the dating points, the language groups' subgrouping adopted for computing the ASJP similarity, and also the ASJP similarity is available in the original paper. We wrote a python program to automatically extract the languages for a given group based on the description given in the original paper. The data for number of languages, calibration date, type of the date, the genetic family, the mode of subsistence (pastoral or agriculture; from the compilation of Hammarström 2010), and the geographic area (based on the continents Eurasia, Africa, Oceania, the Americas) for each language group are given in table B.1.

First, we tested whether the sheer size of the language group (LGS) is related to the calibration dates. The size was determined by counting the number of languages in each language group, using *Ethnologue* (Lewis, Simons and Fennig 2013). A scatter plot with time depth against LGS (on a log-log scale) shows a linear relationship. The regression, shown in figure 10.1, is $r = .81$ and highly significant ($p < 0.001$). The linear relationship is shown by a solid straight regression line. The younger dates are closer to the regression line than the older archaeological dates. figure 10.1 also displays the box plots of each variable along its axis. The box plot of LGS shows three outliers for groups larger than 400, which are farther away from the rest of the dates but not from the regression line. The dotted line is the locally fitted polynomial regression line (LOESS; with degree 2). The LOESS line stays close to the linear regression line confirming that using a linear regression analysis is appropriate. The square root of the variance of the residuals for the LOESS line is also shown as dotted lines on both the sides of the LOESS line.

Although this approach does not require the subgrouping information it is not without problems. The ASJP database often has word lists from dialects of a single language. The ASJP calibration procedure described in ASJP calibration procedure subsection includes all the dialect word lists for a single language identified by its ISO code. Similarly, the LGS variable also counts

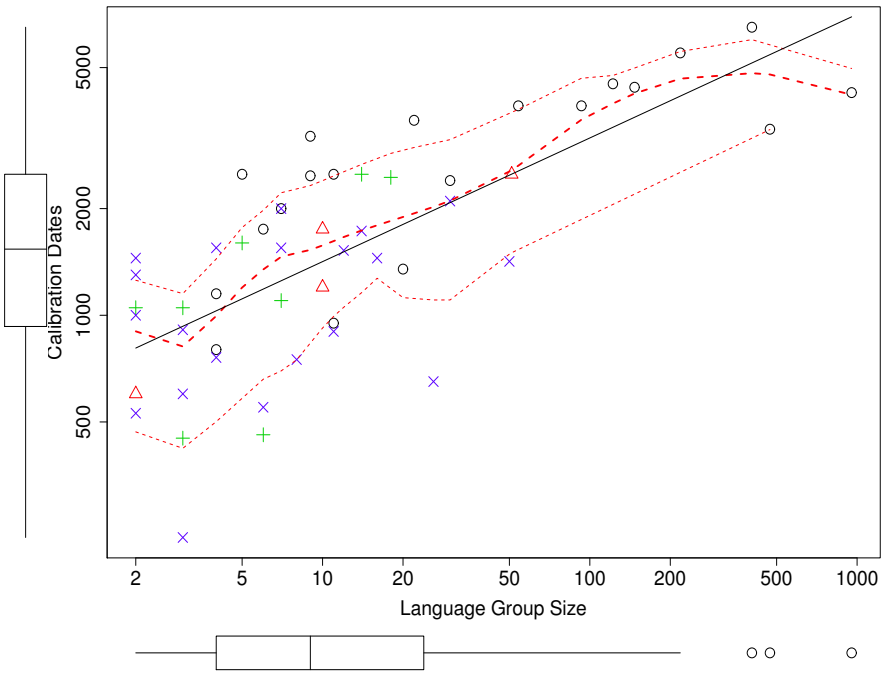


Figure 10.1: Calibration dates against the number of languages in a language group. ○s are archaeological, △s are archaeological and historical, +s are epigraphic and ×s are historical dates.

the total number of available word lists for a language group as its size.[40] Nettle (1999a) summarizes the 'language' vs. 'dialect' judgmental difficulties when adopting language counts from *Ethnologue* for quantifying language diversity (number of languages spoken per unit area). In another work, Nordhoff and Hammarström (2011) use the term 'doculect' to indicate a linguistic variety identified in its descriptive resource. They use this definition to list various language variants in their database *Langdoc*.

In this paper, we follow a different approach which has the following advantages. It requires neither the internal classification information of a language group nor the judgment of language vs. dialect. The approach can use all the available word lists for a language and its dialects identified by a unique ISO 639-3 code. Our approach is described in the next subsection.

[40] We obtain a Pearson's $r = .81$ when LGS variable is counted as the number of languages given in *Ethnologue* (Lewis, Simons and Fennig 2013).

10.2.4 Calibration procedure

In this section, we describe the computation of N-gram diversity and the model selection procedure. The model is run through a battery of tests to check for its robustness. We mix the N-gram model with the ASJP dates to produce a better baseline. Finally, we use the N-gram model to predict the dates for world-wide language groups as given in *Ethnologue*.

10.2.5 N-grams and phonotactic diversity

N-grams are ubiquitous in natural language processing (NLP) and computational linguistics, where they are used in systems ranging from statistical machine translation to speech recognition, but they are relatively unknown in historical linguistics. *N*-grams are defined as a subsequence of length *N* from a sequence of items. The items could be part-of-speech tags, words, morphemes, characters or phonemes. *N*-grams were originally introduced as a probabilistic model for predicting the next linguistic item, given a history of linguistic items (Jurafsky and Martin 2000). The word "oxen" has four letter 1-grams 'o','x','e','n'; three letter 2-grams 'ox', 'xe', 'en'; two letter 3-grams 'oxe', 'xen' and one letter 4-gram 'oxen'. In general, any sequence of length n has $n-N+1$ *N*-grams. The number of *N*-grams can similarly be calculated for a word in an ASJP word list for a given language.

Having introduced *N*-grams, we now define the phonological diversity of a language and show how it can be computed using *N*-grams. Phonological diversity for a language is defined as the size of its phoneme inventory. In a similar fashion, the phonotactic diversity is defined as the total number of possible phoneme combinations in a language. For a language, the 1-gram diversity (computed from a sufficiently long random list of phonetically transcribed words) is the same as phonological diversity. Extending it further, the phonotactic diversity can be computed as the *N*-gram diversity ($N > 1$). Given that the ASJP database (with its wide coverage) is a database of relatively short, 40-item word lists, it needs to be investigated whether the total number of unique phonological segments represented in the 40 item word list can be used as a proxy for the actual phoneme inventory of a language.

Wichmann, Rama and Holman (2011) report a strong positive linear correlation of $r = .61$ between the phoneme inventory sizes for a sample of 392 of the world's languages, from the UPSID database (Maddieson and Precoda 1990b) and the number of phonological segments (which is the same as the 1-gram diversity) represented in word lists for the corresponding

languages in the ASJP database. The mean ratio of the ASJP segment size to the UPSID inventory size is .817 and the standard deviation is .188. Also, there is a small correlation (Pearson's $r = .17$) between the size of the word list, which can vary from 28 to 40, and the number of ASJP phonological segments. This puts us on a solid ground before proceeding to use *N*-grams, extracted from the word lists, for purposes of calibrating dates.

The wide coverage of the ASJP database allows us to provide reasonable relative estimates of the total number of phonological sequences (using ASJPcode) present in the world's languages. Since ASJP modifiers $\sim$ and \$ combine the preceding two or three symbols and there are 41 ASJP symbols in total, the number of theoretically possible phonological sequences is: $41 + 41^2 + 41^3 = 70,643$. But the total number of ASJP sequences varies from 500 to 600 across all languages in the database depending on the criterion for extracting languages from the ASJP database.

The *N*-gram ($N \in [1,5]$) diversity of a language group is defined as the set of all the combined unique phonological segments of length *N* for the languages in the group. One might assume that *N*-grams are not a signature of a language group or, in other words, that *N*-grams do not distinguish unrelated language families from each other. However, it can be empirically established that *N*-grams are more successful in distinguishing unrelated languages from each other than LDND. Wichmann et al. (2010a) devised a measure called *dist*[41] for measuring the efficacy of a lexical similarity measure (in this case LDND vs. LDN) in distinguishing related languages vs. unrelated languages. In a separate experiment, which we will only briefly summarize here, using ASJP data from 49 of the worlds' language families, we employed a 2-gram based measure, *Dice*,[42] for quantifying the distance between the language families and observed that it outperforms LDND in terms of *dist*. This empirical result shows that the set of *N*-grams of a language family is a genetic marker for identifying related vs. unrelated languages.

10.3 Results and Discussion

Objective judgment of shared inheritance of words in related languages becomes increasingly difficult due to the phonological distinctions accumulated over time. We hypothesize that *N*-gram diversity for a language group is a non-decreasing function of time. To verify our hypothesis we check

[41] *Dist* of a family is defined as the difference between intra-family distance and inter-family distances divided by the standard deviation of the inter-family distances.

[42] Between two strings: defined as twice the number of shared bigrams (2-grams) divided by the total number of bigrams.

the nature of relationship between *N*-grams and dates. The last row in figure 10.2 shows the scatterplots of calibration dates (CD; given in table B.1) vs. *N*-grams. The last column of the upper triangular matrix displays significant correlations and the highest correlation between 2-grams and CD. Both 3-grams and 1-grams show a similar correlation with CD whereas, 4-grams and 5-grams show a lower but a similar correlation. Another non-parametric test, Kendall's τ, between the *N*-gram diversity and CD produces a relatively lower but highly significant correlation ($p < 0.001$). The highly significant ρ for different *N*-grams shows that the hypothesis holds for different orders of *N*-grams.

Further, there is a highly significant ρ between *N*-gram diversity and group size, as displayed in figure 10.2. There is a strong correlation between group size and *N*-grams (greater than 0.8 for all *N*). *N*-grams have a highly significant correlation ($p < 0.001$) with each other. Deciding on the optimal value of *N* for the purpose of date calibration is a tricky issue. The LOESS lines for 2- and 3-grams are nearly straight lines compared to the rest of *N*-grams. There needs to be solid evidence for choosing 2- and 3-grams over the rest of *N*-grams. We use the *AIC* measure (Akaike information criterion) coupled with further tests for selecting the appropriate value of *N*. AIC is a relative measure of goodness for model selection. This measure is the negative sum of two components: the number of estimated parameters and the likelihood of the model. The number of parameters is the same across all the *N*-gram models. The lower the AIC, the better is the model. The AIC values for different *N*-grams are given in table 10.1. The values suggest that 2-grams followed by 3-grams are the best fitting models. We employ a generalized linear model (Exponential family – gamma distribution – and log as link function; implementation available as *glm* function in R; R Core Team 2012) with Calibration Dates as the response variable and *N*-grams as predictors.

N	AIC
1	838.05
2	830.52
3	834.84
4	842.84
5	846.08

Table 10.1: The AIC score for each $N-$gram model is displayed in second column. The significance scores for each model compared to the null model are based on a χ^2 test (df $= 50$). All the residual deviance scores are significant at a level of $p < 0.001$.

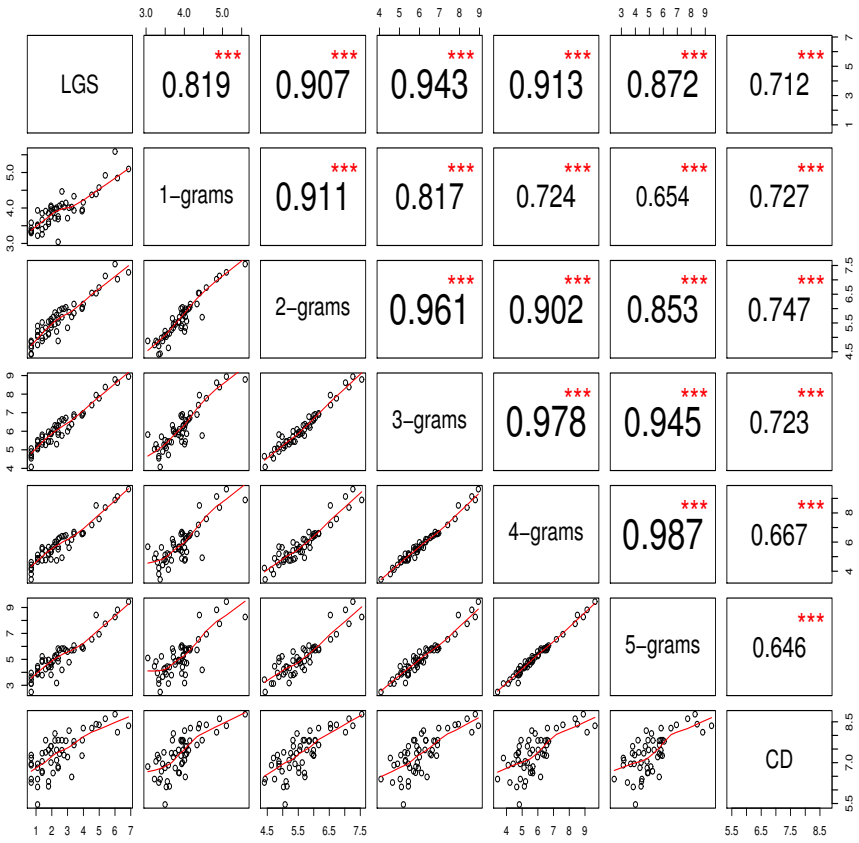


Figure 10.2: Pairwise scatterplot matrix of group size, $N-$gram diversity and date; the lower matrix panels show scatterplots and LOESS lines; the upper matrix panels show Spearman rank correlation (ρ) and level of statistical significance ($\star$). The diagonal panels display variable names. All the plots are on a log-log scale.

Since all calibration dates greater than 2500 BP are archaeological, ASJP tests the significance of the membership of a calibration date in one of the three groups (historical, epigraphic, archaeological) using a one-way analysis of variance (ANOVA). ANOVA tests whether the membership of a date in a group causes bias in the prediction by each N-gram model. The calibration dates are grouped by type of dates, language family, geographical area and mode of subsistence. The data for these groups is available in table B.1. Table 10.2 gives the results of the ANOVA analysis for various groups. The first

column shows the group of the date. The second and third columns show the F-score for algebraic and absolute percent differences for all the N-grams. The fourth column shows the degrees of freedom. The algebraic and absolute percent differences are computed as the percentage of algebraic and absolute residual values to the predicted values.

Group	F, algebraic					df
	1	2	3	4	5	
Type of date	**7.38**	**6.535**	3.217	3.014	3.206	3, 48
Language family	0.61	0.938	1.515	1.441	1.297	16, 35
Geographical area	1.148	1.019	0.533	0.518	0.368	3, 48
Mode of subsistence	2.553	4.152	4.887	2.91	1.988	1, 50
Group	F, absolute					df
	1	2	3	4	5	
Type of date	0.455	1.268	2.357	1.766	1.423	3, 48
Language family	0.572	0.501	1.074	1.049	0.77	16, 35
Geographical area	0.093	0.018	0.677	0.603	0.431	3, 48
Mode of subsistence	0.390	0.272	1.164	0.173	0.04	1, 50

Table 10.2: F-score for algebraic and absolute percentage differences. The significant scores are bold-faced.

Both algebraic and absolute percentages are tested against a significance level of $p < 0.01$. The test suggests that the predicted dates of 1-grams and 2-grams are biased in terms of type of the dates. The test suggests that the bias is with respect to archaeological class of dates. All the other values are non-significant and suggest that there is no difference across the groups. Thus, the ANOVA analysis suggests that the 3-gram dates are more robust than 2-gram dates and are unbiased with respect to the groups.

We now test the validity of the assumptions of the regression analysis through the standard diagnostic plots, given in section B.2 – figures B.1, B.2, B.3, B.4, and B.5. The diagnostic plots of 3-gram model in figure B.3 suggest that there has been no violation in the assumptions of regression analysis. The scatterplot between the predicted values and the residuals do not show any pattern. The residuals are normally distributed and the plot suggests that Dardic and Southwest Tungusic groups are the most deviating points. The normality assumption of the residuals is further tested through a Kolmogorov-Smirnov test (KST). KST tests against the null hypothesis that the residuals are distributed normally under a significance criterion of $p < 0.01$. The test gives a $p = .86$ suggesting that we can retain the null

hypothesis of normality. The ASJP dates for Dardic is underestimated by 90% and overestimated for Southwest Tungusic by 72%. The 3-gram dates for Dardic and Southwest Tungusic are 1743 BP and 1085 BP, respectively. It is not clear why there is such a huge discrepancy for these dates. The influential and leverage points are identified in subplot 3 (in figure B.3). The diagnostic plot does not suggest any influential points whereas there seems to be atleast five high leverage points in the plot. The leverage points are identified as Benue-Congo, Eastern Malayo-Polynesian, Ga-Dangme, Indo-European and Malayo-Polynesian. All these points are archaeological and exceed a time depth of 3500 years (except for Ga-Dangme which is both archaeological and historical and only 600 years old). As a matter of fact, the absolute percentage difference with respect to ASJP dates are as follows: $-32, +12, -37, -26$ and -41.

Summarizing the regression analysis, there is a strong correlation of .723 between the logarithm of 3-gram diversity and the calibration dates. We tested the assumptions of regression analysis and find that they are not violated. The 3-gram diversity reflects the net phonotactic diversity accumulated or lost in a language group over time. The predictions of all the N-gram models and the respective calibration date are presented in figure 10.3.

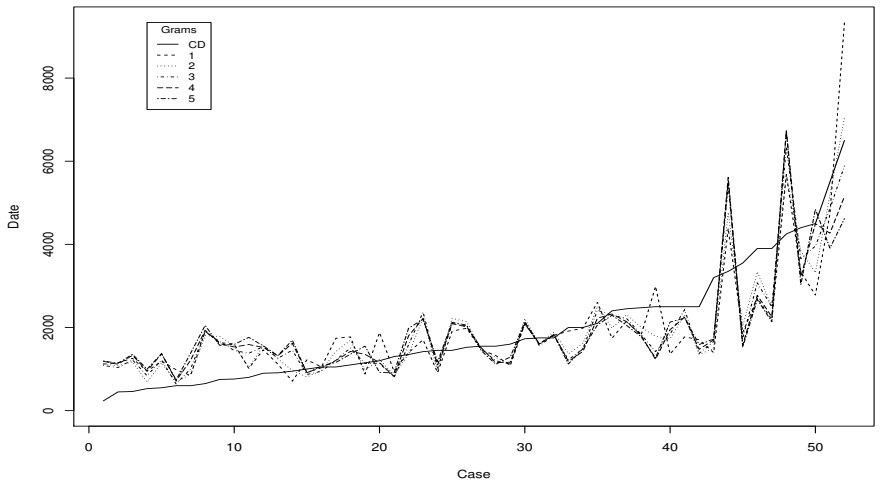


Figure 10.3: Comparing predicted dates for various n-grams

The current findings can be explained in the terms of the basic model of language change described in Nettle 1999a. In this model, languages diverge through imperfect replication of linguistic items amplified by selectional

pressure and geographic isolation. Selectional pressures, namely social and functional selection, operate in the selection of the language variants generated through imperfect learning and the learner's performance in this language. 3-grams are a proxy for phonotactic diversity. The difference in phonotactic diversity between two languages represents the net result of phonological erosion, morphological expansion and fusion the language has undergone since its divergence from its most recent shared ancestor. The correlation between 3-grams and time depth is just the reflection of this strong relation with net phonotactic diversity.

Since ASJP dates and 3-gram dates use different information from the same database, it would be interesting to see how the mixture of the predictions of the two models fare against the calibration dates. Each ASJP date is combined with a 3-gram date using the following formula:

$$COD = k * ASJPD + (1-k) * NGD \tag{11}$$

where $0 < k < 1$, ASJPD is a ASJP date, NGD is either 2-gram or 3-gram dates and, COD is a combined date. For a value of k, ranging from 0 to 1, the value of ρ between COD and calibration dates is plotted in figure 10.4. The horizontal axis displays the scaled k ranging from 0 to 100. Figure 10.4 shows that there is a modest, but steady increase in the correlation when ASJP dates are combined with 3-gram dates. The correlation increases until 40% and then remains stable from 40% to 60%. Both 2-grams and 3-grams show the same trend. This indicates that a combination of the predictions indeed works better than individual models for the uncalibrated language families of the world. The optimal combination for 3-grams is obtained at $k = .59$.

The effect of mixing of 3-gram dates with ASJP dates is tested in table 10.3. Table 10.3 gives a comparison of ASJP dates, 3-gram dates, and combined dates in terms of: sum and mean of absolute discrepancy, number of languages off by 50% and 100%, and ρ. The ASJP analysis gave an upper bound of 29% on the expected discrepancy between ASJP dates and the true dates for different language groups. We observe that the average of the absolute percentage discrepancy of combined dates (18%) falls within the range of ASJP discrepancy. Clearly combined dates outperforms both the ASJP and 3-gram model's methods. 3-gram dates have the advantage that they neither requires subgrouping information nor the distinction between 'language' and 'dialect' but does not have the same ρ as ASJP dates. Combined dates performs the best but is the most complicated and has the disadvantages of ASJP dating.

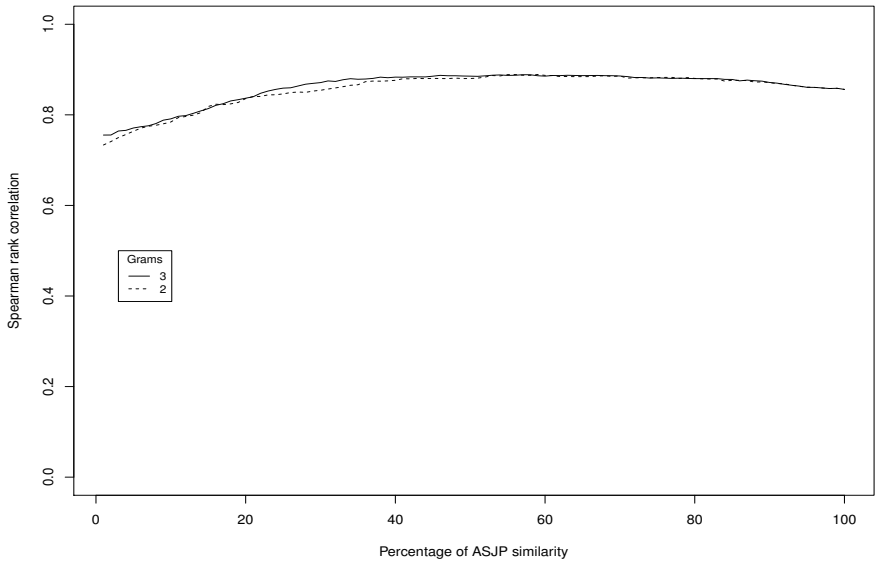


Figure 10.4: Combining ASJP with 2−grams and 3−grams: The ASJP dates are combined with 2−gram dates and 3−gram dates in different proportions ranging from 1% to 100% at an interval of 1.

Measurement	ASJP	3−grams	combined
Sum of absolute discrepancy	1523	1815	927
Mean of absolute discrepancy	29	34	18
Off by 50%	5	13	2
Off by 100%	1	1	0
Spearman's ρ	.86	.72	.89

Table 10.3: A comparison of different dating methods

10.3.1 Worldwide date predictions

Finally, we predict time depths for the world's language families, as given in *Ethnologue*, using the 3-gram model. A combined date is given through Eq. 11. Both the predicted and the combined dates are given in tables B.2–B.6 (section B.3). Each table presents the dates for all language families belonging to a geographical area – as defined in section 10.2. The first column of each table shows the name of a language family and its subgroups (if any). For

each language family, a subgroup and its further internal classifications are indented. For the sake of comparison, we give dates only for those families and subgroups given by ASJP (Holman et al. 2011). The second column in each table shows the number of languages for a subgroup. The third and fourth columns show the ASJP dates and the 3-gram predicted dates. The fifth column shows the combined date, computed using Eq. 11. Whenever the ASJP date is missing for a language group we did not compute a combined date.

We now comment on the level of agreement found between ASJP dates and 3-gram dates in tables B.2–B.6 and try to explain the differences in terms of known linguistic or geographic factors. Except for Khoisan, the ASJP dates as well as 3-gram dates are quite similar. The language families Afro-Asiatic, Nilo-Saharan, and Niger-Congo are quite old and here the dates are similar. There is an enormous difference between the two dates for Khoisan. ASJP predicts $14,500$ years as the time depth of Khoisan family whereas 3-grams predict a shallower date ($1,863$ years). This huge disagreement could be attributed to the many-to-one mapping of click consonants by ASJP code. Additionally, ASJP (Holman et al. 2011) noted that some of the family classifications given in *Ethnologue* are controversial. Such a huge time gap could be a result of a lack of consensus in the general definition of a language family.

There is a relatively minor difference between the dates in Holman et al. (2011) and 3-gram dates for the well-established language families of Eurasia such as Austro-Asiatic, Dravidian, Indo-European, Sino-Tibetan, and Uralic (table B.3). Both models predict similar dates for Eurasian language families. The dates for languages of Pacific area is given in table B.4. For Austronesian, a large language family (974 languages) in the Pacific area, the ASJP and 3-gram dates are $3,633$ and $6,455$ years, respectively. The combined date of Austronesian family is $4,790$ years which is fairly close to the age given by Greenhill, Drummond and Gray (2010), $5,100$ years.

3-gram dates and ASJP dates differ greatly for nearly all the language families of North America (table B.5). For instance, ASJP Holman et al. (2011) predict a time depth of $5,954$ years for Algic whereas 3-grams predict $3,236$ years. The 3-gram dates and ASJP dates differ by a few decades for the Mixe-Zoque and Mayan families, which are spoken in Middle America. A similar kind of difference is evident for a majority of South American languages (table B.6). In summary, the ASJP and 3-gram dates' differences cannot be explained in terms of geographical areas. A huge gap between ASJP and 3-gram dates, such as Khoisan, might be a potential signal for a phantom phylogeny.

10.4 Conclusion

In this paper we replicated the ASJP consortium's process of extracting data representative of 52 language groups for the use of calibrating linguistic chronologies. We proposed *N*-gram diversity as a measure of phonotactic diversity and found that 3-gram diversity had a significant correlation of 0.72 with calibration dates. The most important finding was that a combination of ASJP lexical similarity and 3-gram diversity, currently, is the best baseline for predicting the time depths for a language family. Finally, time depths for worldwide language families were predicted and combined with ASJP dates. The new dates are provided in section B.3.

11 LINGUISTIC LANDSCAPING OF SOUTH ASIA

Borin Lars, Anju Saxena, Taraka Rama, and Bernard Comrie 2014. Linguistic landscaping of South Asia using digital language resources: Genetic vs. areal linguistics. *Proceedings of The International Conference on Language Resources and Evaluation*. 3137–3144.

Abstract

Like many other research fields, linguistics is entering the age of big data. We are now at a point where it is possible to see how new research questions can be formulated – and old research questions addressed from a new angle or established results verified – on the basis of exhaustive collections of data, rather than small, carefully selected samples. For example, South Asia is often mentioned in the literature as a classic example of a linguistic area, but there is no systematic, empirical study substantiating this claim. Examination of genealogical and areal relationships among South Asian languages requires a large-scale quantitative and qualitative comparative study, encompassing more than one language family. Further, such a study cannot be conducted manually, but needs to draw on extensive digitized language resources and state-of-the-art computational tools. We present some preliminary results of our large-scale investigation of the genealogical and areal relationships among the languages of this region, based on the linguistic descriptions available in the 19 tomes of Grierson's monumental *Linguistic Survey of India* Grierson (1927), which is currently being digitized with the aim of turning the linguistic information in the LSI into a digital language resource suitable for a broad array of linguistic investigations.

11.1 Introduction

Like many other research fields, linguistics is entering the age of big data. The modern digital world and the mass digitization of historical documents together provide unprecedented opportunities to linguistics and other disciplines relying on text and speech as primary research data. However, this development comes with considerable methodological challenges. We are now at a point where it is possible to see how new research questions can be formulated – and old research questions addressed from a new angle or established results verified – on the basis of exhaustive collections of data, rather than small, carefully selected samples, but where a methodology has not yet established itself, and where serious studies have hardly been conducted at all.

For example, comprehensive, large-scale quantitative and qualitative studies are essential in order to get a deeper understanding of areal linguistics. South Asia[43] is often mentioned in the literature as a classic example of a linguistic area. There is, however, no systematic, empirical study of South Asian languages to substantiate this claim. In order to critically evaluate South Asia as a linguistic area, a systematic examination of a set of linguistic features in a wide range of South Asian languages is essential.

South Asian languages belong to four major language families: Indo-European (>Indo-Aryan), Dravidian, Austroasiatic (>Mon-Khmer and Munda) and Sino-Tibetan (>Tibeto-Burman). There are also some small families (e.g., in the Andaman Islands), some language isolates (e.g., Burushaski and Nihali), and some unclassified languages.

Throughout history multilingualism has been the norm in the area. There are signs of language contact between Vedic Sanskrit and Dravidian languages in the Rig Veda, the oldest text found in India. It has been claimed that this long-lasting contact situation has made the languages of this region more similar in some respects to each other than they are to their genealogically related languages spoken outside this region, and that consequently South Asia should be seen as a linguistic area (e.g., (Emeneau 1956; Masica 1976; Kachru, Kachru and Sridhar 2008), and others). However, systematic investigations of this claim have been few and somewhat spotty, mostly relying on data from a few major Indo-Aryan and Dravidian languages (see Ebert 2006 for a critique). The approach of Subbarao 2008 is representative: Linguistic features (most of them from Emeneau 1956) are illustrated with single – 'cherry-picked' – linguistic examples, and different

[43] Although South Asia is defined variously in the literature, in linguistic works this area is usually considered to comprise the seven countries Bangladesh, Bhutan, India, the Maldives, Nepal, Pakistan, and Sri Lanka.

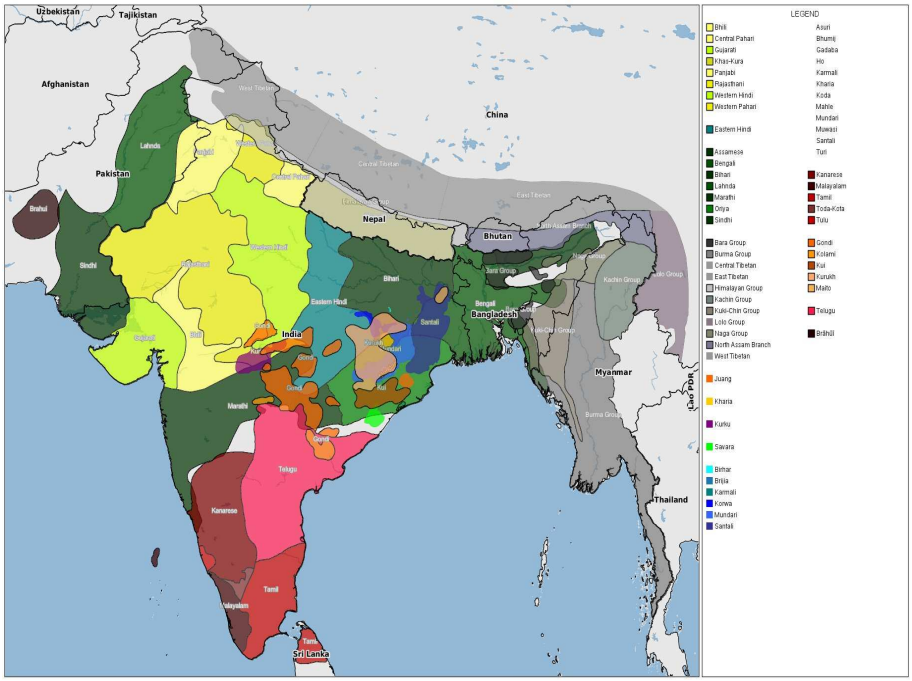
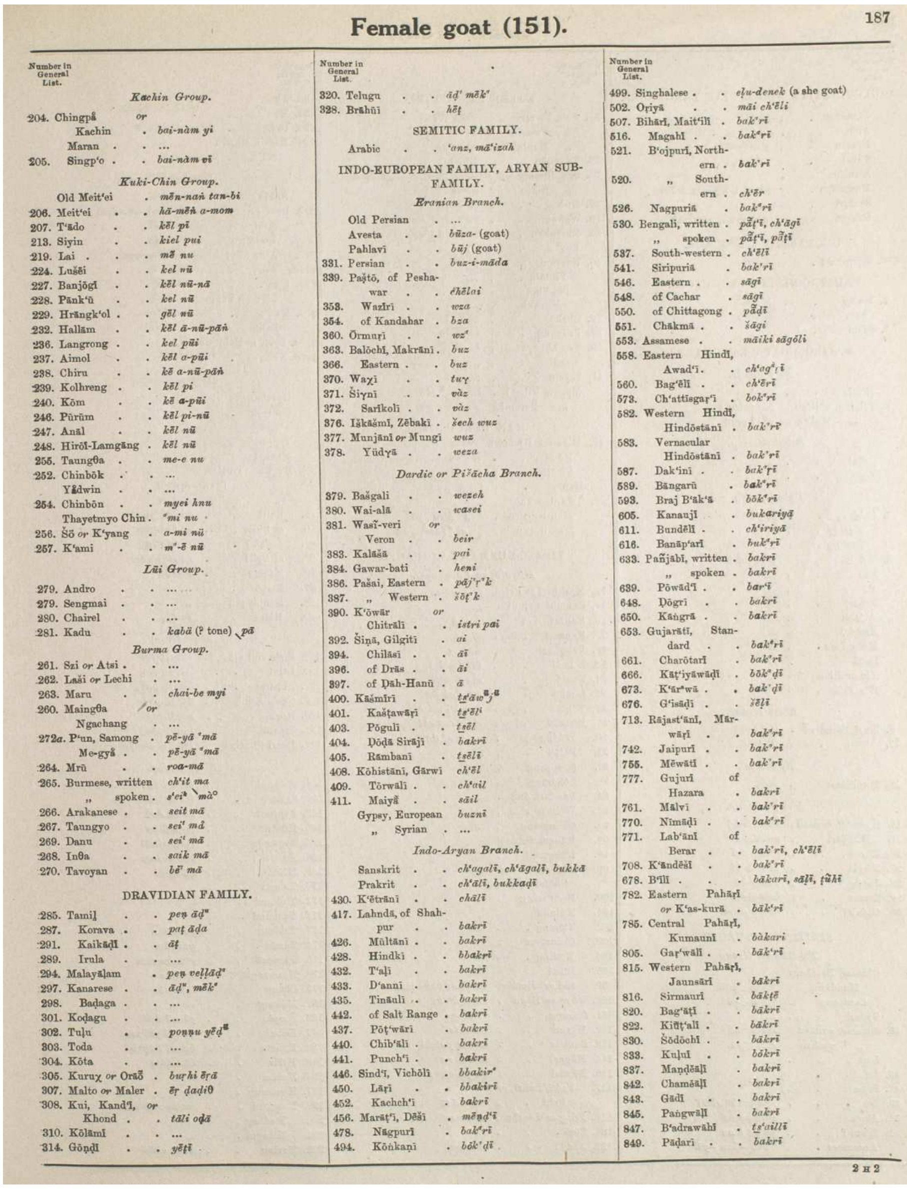


Figure 11.1: Map of four major South Asian language families (from `http://llmap.org`)

languages are used to illustrate different linguistic features. This is understable at one level: One would like to include as many languages as possible in a study, and doing the work manually puts severe restrictions as to how many languages and/or features one can handle.

In order to critically evaluate the notion of South Asia as a linguistic area, we need to know the spread and extent of a linguistic feature across space and language families. Further, the internal sub-grouping of all the South Asian language families remains unclear. E.g., Asher (2008) problematizes the current internal subgrouping of the Indo-Aryan language family, as the proposed subgroups correlate highly with their geographical distribution.[44] The focus in works on the internal relationships is on one family at a time, e.g., Turner 1964, Bloch 1954, Cardona and Jain 2003 on Indo-Aryan; Burrow and Emeneau 1984, Krishnamurti 2003 on Dravidian; Matisoff 2003

[44]Correlation with geographical distribution is not *inherently* a problem, of course. If a number of speech communities have migrated minimally since the split-up of their ancestor language, then one would expect geography and genealogy to correlate. But this then makes it hazardous to attempt to distinguish between effects of geography and geneaology on the basis of language-internal evidence alone.



Female goat (151). 187

Number in General List.

Kachin Group.

204. Chingpå or Kachin . *bai-nàm yi*
Maran . ...
205. Singp'o . *bai-nàm vī*

Kuki-Chin Group.

Old Meit'ei . *mēn-naṅ tan-bi*
206. Meit'ei . *hā-mēṅ a-mom*
207. T'ādo . *kēl pī*
213. Siyin . *kiel pui*
219. Lai . *mē nu*
224. Lušēi . *kel nū*
227. Banjōgī . *kēl nū-nā*
228. Pānk'ū . *kel nū*
229. Hrāngk'ol . *gēl nū*
232. Hallām . *kēl ā-nū-pāṅ*
236. Langrong . *kel pūi*
237. Aimol . *kēl a-pūi*
238. Chiru . *kē a-nū-pāṅ*
239. Kolhreng . *kēl pi*
240. Kōm . *kē a-pūi*
246. Pūrūm . *kēl pi-nū*
247. Anāl . *kēl nū*
248. Hirōī-Lamgāng . *kēl nū*
255. Taungθa . *me-e nu*
252. Chinbōk . ...
Yådwin . ...
254. Chinbōn . *myei hnu*
Thayetmyo Chin . *'mi nu*
256. Šō or K'yang . *a-mi nū*
257. K'ami . *m'-ē nū*

Lūi Group.

279. Andro . ...
279. Sengmai . ...
280. Chairel . ...
281. Kadu . *kabä* (? tone) *pā*

Burma Group.

261. Szi or Atsi . ...
262. Laši or Lechi . ...
263. Maru . *chai-be myi*
260. Maingθa or Ngachang . ...
272a. P'un, Samong . *pē-yā 'mā*
Me-gyå . *pē-yā 'mā*
264. Mrū . *roa-mā*
265. Burmese, written *ch'it ma*
" spoken . *s'ei mà°*
266. Arakanese . *seit mā*
267. Taungyo . *sei' mâ*
269. Danu . *sei' mā*
268. Inθa . *saik mā*
270. Tavoyan . *bē' mā*

DRAVIDIAN FAMILY.

285. Tamiḻ . *peṇ āḍ*
287. Korava . *paṭ āḍa*
291. Kaikāḍī . *āṭ*
289. Irula . ...
294. Malayāḷam . *peṇ veḷḷāḍ*
297. Kanarese . *āḍ, mēk'*
298. Baḍaga . ...
301. Koḍagu . ...
302. Tuḷu . *poṇṇu yēḍ*
303. Toda . ...
304. Kōta . ...
305. Kuruχ or Orāō . *buṛhi ēṛā*
307. Malto or Maler . *ēṛ ḍaḍiθ*
308. Kui, Kand'ī, or Khond . *tāli oḍā*
310. Kōlāmī . ...
314. Gōṇḍī . *yēṭī*
320. Telugu . *āḍ' mēk'*
328. Brāhūī . *hēṭ*

SEMITIC FAMILY.

Arabic . *'anz, mā'izah*

INDO-EUROPEAN FAMILY, ARYAN SUB-FAMILY.

Eranian Branch.

Old Persian . ...
Avesta . *būza-* (goat)
Pahlavī . *būj* (goat)
331. Persian . *buz-i-māda*
339. Paštō, of Peshawar . *ċhēlai*
353. Wazīrī . *wza*
354. of Kandahar . *bza*
360. Ōrmuṛī . *wz'*
363. Balōchī, Makrānī . *buz*
366. Eastern . *buz*
370. Waχī . *tuγ*
371. Šiγnī . *vùz*
372. Sarīkolī . *vùz*
376. Iškāšmī, Zēbakī . *šech wuz*
377. Munjānī or Mungī *wuz*
378. Yūdγā . *weza*

Dardic or Piśācha Branch.

379. Bašgali . *wezeh*
380. Wai-alā . *wasei*
381. Wasī-veri or Veron . *beir*
383. Kalāšā . *pai*
384. Gawar-bati . *heni*
386. Pašai, Eastern . *pāj'ṛ'k*
387. " Western . *žōṭ'k*
390. K'ōwār or Chitrālī . *istri pai*
392. Šiṇā, Gilgitī . *ai*
394. Chilāsī . *āī*
396. of Drās . *āi*
397. of Ḍāh-Hanū . *ā*
400. Kāšmīrī . *ts'āwj*
401. Kašṭawāṛī . *ts'ēl*
403. Pōgulī . *tsēl*
404. Ḍōḍā Sirājī . *bakrī*
405. Rāmbanī . *tsēlī*
408. Kōhistānī, Gārwī *ch'ēl*
409. Tōrwālī . *ch'ail*
411. Maiyā̃ . *sāil*
Gypsy, European *buznī*
" Syrian . ...

Indo-Aryan Branch.

Sanskrit . *ch'agalī, ch'āgalī, bukkā*
Prakrit . *ch'ālī, bukkaḍī*
430. K'ētrānī . *chālī*
417. Lahndā, of Shahpur . *bakrī*
426. Mūltānī . *bakrī*
428. Hindkī . *bbakrī*
432. T'aḷī . *bakrī*
433. D'annī . *bakrī*
435. Tināulī . *bakrī*
442. of Salt Range . *bakrī*
437. Pōṭ'wārī . *bakrī*
440. Chib'ālī . *bakrī*
441. Punch'ī . *bakrī*
446. Sind'ī, Vichōlī . *bbakir'*
450. Lāṛī . *bbakirī*
452. Kachch'ī . *bakrī*
456. Marāṭ'ī, Dēšī . *mēṇḍ'ī*
478. Nāgpurī . *bak'rī*
494. Kōṅkaṇī . *bōk'ḍī*
499. Singhalese . *eḷu-denek* (a she goat)
502. Oṛiyā . *māi ch'ēli*
507. Bihārī, Mait'ilī . *bak'rī*
516. Magahī . *bak'rī*
521. B'ojpurī, Northern . *bak'rī*
520. " Southern . *ch'ēr*
526. Nagpuriā . *bak'rī*
530. Bengali, written . *pā̃ṭ'ī, ch'āgī*
" spoken . *pā̃ṭ'ī, pā̃ṭī*
537. South-western . *ch'ēlī*
541. Siripuriā . *bak'rī*
546. Eastern . *sāgī*
548. of Cachar . *sāgī*
550. of Chittagong . *pā̃ḍī*
551. Chākmā . *šāgi*
553. Assamese . *māiki sāgōli*
558. Eastern Hindī, Awad'ī . *ch'ag'ṛī*
560. Bag'ēlī . *ch'ērī*
573. Ch'attīsgaṛ'ī . *bok'rī*
582. Western Hindī, Hindōstānī . *bak'rī*
583. Vernacular Hindōstānī . *bak'rī*
587. Dak'inī . *bak'ṛī*
589. Bāngarū . *bak'rī*
593. Braj B'āk'ā . *bōk'rī*
605. Kanaujī . *bukariyā*
611. Bundēlī . *ch'iriyā*
616. Banāp'arī . *buk'rī*
633. Pañjābī, written . *bakrī*
" spoken . *bakrī*
639. Pōwād'ī . *bar'ī*
648. Ḍōgrī . *bakrī*
650. Kāṅgrā . *bakrī*
653. Gujarātī, Standard . *bak'rī*
661. Charōtarī . *bak'rī*
666. Kāṭ'iyāwāḍī . *bōk'ḍī*
673. K'ār'wā . *bak'ḍī*
676. G'isāḍī . *šēḷī*
713. Rājast'ānī, Mārwāṛī . *bak'rī*
742. Jaipurī . *bak'rī*
755. Mēwātī . *bak'rī*
777. Gujurī of Hazara . *bakrī*
761. Mālvī . *bak'rī*
770. Nīmāḍī . *bak'rī*
771. Lab'ānī of Berar . *bak'rī, ch'ēlī*
708. K'āndēšī . *bak'rī*
678. B'īlī . *bākarī, sāḷī, ṭūhī*
782. Eastern Pahāṛī or K'as-kurā . *bāk'rī*
785. Central Pahāṛī, Kumaunī . *bàkari*
805. Gaṛ'wālī . *bāk'rī*
815. Western Pahāṛī, Jaunsārī . *bākrī*
816. Sirmaurī . *bākṭē*
820. Bag'āṭī . *bākrī*
822. Kiūṭ'alī . *bākrī*
830. Šōdōchī . *bākrī*
833. Kuḷuī . *bōkrī*
837. Maṇḍēāḷī . *bakrī*
842. Chamēāḷī . *bakrī*
843. Gādī . *bakrī*
845. Paṅgwāḷī . *bakrī*
847. B'adrawāhī . *ts'aillī*
849. Pāḍarī . *bakrī*

2 H 2

Figure 11.2: The LSI comparative vocabulary.

and Thurgood and LaPolla 2003 on Sino-Tibetan. However, given the claims about South Asia as a linguistic area, it would be prudent to always have an eye open for contact influences from other families, since these might vitiate aspects of a purely family-internal investigation.

The map in figure 11.1 shows the geographical extent of the four major South Asian language families (Austroasiatic, Dravidian, Indo-Aryan, and Sino-Tibetan), including overlap between languages belonging to different families. Three prominent cases of such overlap are Gondi (Dravidian)–Marathi (Indo-Aryan), Brahui (Dravidian)–Sindhi (Indo-Aryan), and Santali (Munda)–Bengali (Indo-Aryan).

11.2 Towards a language resource from Grierson's LSI

Examination of genealogical and areal relationships among South Asian languages requires a large-scale comparative study, encompassing more than one language family. Further, such a study cannot be conducted manually, but needs to draw on extensive digitized language resources and state-of-the-art computational tools. In this paper we will present some preliminary results of our large-scale investigation of the genealogical and areal relationships among the languages of this region, based on the linguistic material available in Grierson's *Linguistic Survey of India* (LSI; Grierson 1927), which is currently being digitized with the aim of turning the linguistic information in the LSI into a digital language resource, a database suitable for a broad array of linguistic investigations, which will be made freely available under an open-content license.[45]

The LSI still remains the most complete single source on South Asian languages. Its 19 tomes (9500 pages) cover 723 linguistic varieties representing major language families and some unclassified languages, of almost the whole of nineteenth-century British-controlled India (modern India, Pakistan, Bangladesh, and parts of Burma). For each major variety it provides (1) a grammatical sketch (including a description of the sound system); (2) a core word list; and (3) texts (including a translation of the Parable of the Prodigal Son). The core word lists which accompany the language descriptions are collected in a separate volume (Volume 1, Part 2: *Comparative vocabulary*; see figure 11.2).

Each list has a total of 168 entries (concepts). The concepts in the comparative vocabulary cover a broad spectrum consisting of body parts, domestic animals, personal pronouns, numerals, and astronomical objects.

There is some overlap with other concept lists used in language classification: First, 38 of the concepts are also found in the shorter (100-item) version of the so-called *Swadesh lists*, core vocabulary lists

[45] See, e.g., the IDS wordlists (Borin, Comrie and Saxena 2013) produced in our ongoing project, available under a CC-BY license from `http://spraakbanken.gu.se/eng/research/digital-areal-linguistics/word-lists`.

originally devised by the American linguist Morris Swadesh (Swadesh 1950, 1952, 1955) specifically for the purpose of inferring genealogical relationships among languages. Further, 76 of the items are found in an extended Swadesh list used by us in earlier genealogical investigations of Tibeto-Burman languages of the Indian Himalayas (e.g. Saxena 2011; Saxena and Borin 2011, 2013). Similarly, 34 LSI vocabulary items are present in the Leipzig-Jakarta list, a 100-item list of word senses claimed to be highly resistant to borrowing (Haspelmath and Tadmor 2009b).

Thus, the LSI comparative vocabulary clearly has one part that can be used in investigating genetic connections among the languages, but also another part – at least half of the entries – which we hypothesize could be used to find areal influences.

Family	# varieties
Austro-Asiatic	12
Dravidian	19
Indo-Aryan	96
Sino-Tibetan	141

Table 11.1: Major South Asian family languages in the LSI comparative vocabulary

11.3 Some preliminary experiments

In this paper, we focus on the data extracted from the comparative vocabulary. All in all, the LSI offers core vocabulary lists for more than 250 language varieties from the four main South Asian language families (see table 11.1).

The wide range of languages and the number and type of concepts present in this language resource allow us address the issue of genealogical vs. areal factors using computational methods (Wichmann 2008). These computational methods take a set of vocabulary lists as input, and yield a distance matrix between the vocabulary lists as output. The distance matrix may subsequently be used as an input to a phylogenetic program to infer a classification tree for the set of languages.

The distance matrix is computed through the application of a variant of Levenshtein distance (Levenshtein 1966), LDND (Levenshtein Distance Normalized Double; Wichmann et al. 2010a). The matrix is then given to a Neighbor-Joining program (Saitou and Nei 1987) to yield an unrooted tree. For our data, this cross-family tree groups the languages into distinct clusters corresponding to the recognized language families.

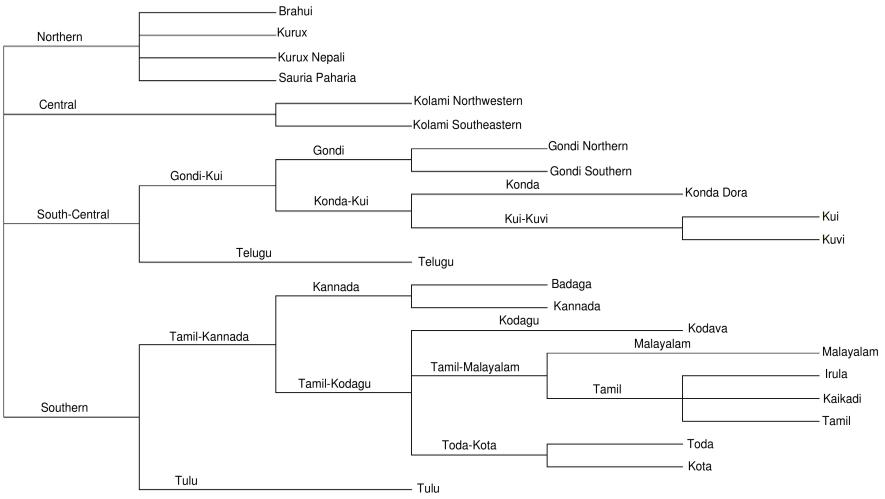


Figure 11.3: The Dravidian family tree according to *Ethnologue*

Subsequently, we evaluate the overall accuracy of classification in each language family through a comparison of a family's distance matrix with its gold-standard family tree, extracted from the *Ethnologue* (Lewis, Simons and Fennig 2013).[46] An example of such a tree, for the Dravidian language family, is shown in figure 11.3.

One way to compare the distance matrix with the family tree is to compute the correlation between pair-wise language distances and the pair-wise branch lengths (as read off from the tree). However, it is not obvious how to best compute the distances from the family tree.

The most straightforward method would be the *raw branch length*, or simply the number of nodes encountered in the shortest path lying from a language A to a language B. The left-most plot in figure 11.4 shows the agreement between LDND distance and pair-wise raw branch length distance for the Dravidian language family. The fit is not particularly good. There could be more than one reason for this, of course, but the naïve raw branch length method is certainly a strong suspect, given the standard assumption that observed language change – and consequently the distance between related language varieties – should be a function of time depth.

If we then assume that the horizontal dimension in figure 11.3 reflects time depth, all terminal nodes in the family tree should lie at the same distance

[46]The Ethnologue is not above reproach as a gold standard, but at present there is hardly a better comprehensive source of language family information covering such a broad range of languages.

from the root node, since they all represent present-day languages with an equally long history of descent from the proto-language. Typical language family trees are 'unbalanced', with a different number of intermediate nodes on different branches, as the Dravidian tree in figure 11.3, where the Northern Dravidian language Brahui is much closer to the root than Southern Dravidian Malayalam.

We propose to balance the Ethnologue trees in the following way: Each node is weighted according to its depth from the root node, with the maximal depth (the terminal node furthest away from the root) always set to be 1. For instance, the Northern Dravidian node would get a weight of 0.5 (1/2) as opposed to the Southern Dravidian node which is assigned a weight of 0.17 (1/6). The right-most plot in figure 11.4 shows the agreement when the branch lengths are computed from the balanced Dravidian family tree.

The agreement between the distance matrix and the branch length matrix can be computed using Kendall's τ, a rank-based correlation measure highly suited for this purpose, since it takes ties in ranks into account.[47] There is a significant ($p < 0.001$) improvement of τ from the use of raw branch length to balanced tree branch length.[48]

Another way of computing the agreement between the gold-standard tree and the LDND distances is offered by a modified version of Goodman-Kruskal's γ (Goodman and Kruskal 1954). This score compares language triplets wrt whether both measures show the same distance relations among the languages in a triplet or not (ignoring ties). γ lies in the range [-1, 1] where a score of -1 indicates total disagreement and a score of 1 indicates perfect agreement.

We computed γ for the four major South Asian language families and found that LDND agrees with the balanced branch length score better than the raw branch length score (see table 11.2).

Family	γ raw	γ balanced
Austroasiatic	0.6379	0.7766
Dravidian	0.4235	0.748
Indo-Aryan	0.479	0.6517
Sino-Tibetan	0.3817	0.5416

Table 11.2: Goodman-Kruskal's γ for the four major South Asian language families

[47] Because of the tree topology, there will normally be more than one language pair with the same distance between them.

[48] The statistical significance of the scores is evaluated through a Mantel's test (Mantel 1967), a permutation test which computes the significance of a test statistic by permuting the rows of a matrix and recomputing the τ score. This procedure is repeated 1000 times.

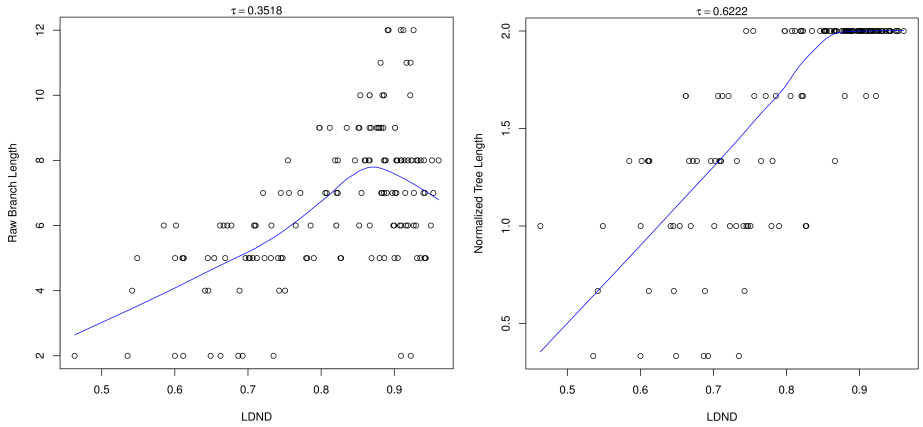


Figure 11.4: Distance matrix fit to Dravidian raw tree (left) and balanced tree (right) shown by locally fitted regression lines.

11.4 Discussion, conclusions and outlook

The experiments described above have shown how LDND distance calculations on the LSI comparative vocabulary recover both inter-family and intra-family genealogical relations for the four major South Asian language families. However, no indications of areal phenomena could be seen using this method on the LSI comparative vocabulary. The phylogenetic trees built from the distance matrix cluster related languages together, whereas no cross-family areal clusters emerge. The correlation between LDND distances and balanced family tree distances was high for both measures used, whereas with a strong areal component, a lower correlation would have been expected.

There are several conceivable reasons for this, e.g.: (1) Contrary to our expectations (see section 11.2), the LSI comparative vocabulary is the 'wrong' vocabulary for uncovering language contact; (2) the comparison method chosen (LDND) is not suitable for this problem; or (3) we need to look at other parts of the language than vocabulary in order to establish areal connections.

Unfortunately, the LSI cannot help us with (1). As for (2), we have made some experiments using a more linguistically informed semi-automatic vocabulary comparison showing very encouraging results (Saxena 2011; Saxena and Borin 2011, 2013). Future research will show if this methodology will scale up sufficiently to deal with the LSI comparative vocabulary. Finally, with respect to (3), we are in the process of extracting the information on the various grammatical features found in the LSI grammar sketches into a rich typological database, which will hopefully provide us with a firmer basis for investigating areal and micro-areal phenomena in South Asia.

12 STRING SIMILARITY MEASURES FOR LANGUAGE CLASSIFICATION

Rama, Taraka and Lars Borin 2015. Comparative Evaluation of String Similarity Measures for Automatic Language Classification. Ján Mačutek and George K. Mikros (eds), *Sequences in language and text*, 203–231. Walter de Gruyter.

12.1 Introduction

Historical linguistics, the oldest branch of modern linguistics, deals with language-relatedness and language change across space and time. Historical linguists apply the widely-tested comparative method (Durie and Ross 1996) to establish relationships between languages to posit a *language family* and to reconstruct the proto-language for a language family.[49] Although historical linguistics has parallel origins with biology (Atkinson and Gray 2005), unlike the biologists, mainstream historical linguists have seldom been enthusiastic about using quantitative methods for the discovery of language relationships or investigating the structure of a language family, except for Kroeber and Chrétien (1937) and Ellegård (1959). A short period of enthusiastic application of quantitative methods initiated by Swadesh (1950) ended with the heavy criticism levelled against it by Bergsland and Vogt (1962). The field of computational historical linguistics did not receive much attention again until the beginning of the 1990s, with the exception of two noteworthy doctoral dissertations, by Sankoff (1969) and Embleton (1986).

In traditional lexicostatistics, as introduced by Swadesh (1952), distances between languages are based on human expert *cognacy judgments* of items in

[49]The Indo-European family is a classical case of the successful application of comparative method which establishes a tree relationship between some of the most widely spoken languages in the world.

standardized word lists, e.g., the Swadesh lists (Swadesh 1955). In the terminology of historical linguistics, *cognates* are related words across languages that can be traced directly back to the proto-language. Cognates are identified through regular sound correspondences. Sometimes cognates have similar surface form and related meanings. Examples of such revealing kind of cognates are: English ∼ German *night* ∼ *Nacht* 'night' and *hound* ∼ *Hund* 'dog'. If a word has undergone many changes then the relatedness is not obvious from visual inspection and one needs to look into the history of the word to exactly understand the sound changes which resulted in the synchronic form. For instance, the English ∼ Hindi *wheel* ∼ *chakra* 'wheel' are cognates and can be traced back to the proto-Indo-European root k^wek^wlo-.

Recently, some researchers have turned to approaches more amenable to automation, hoping that large-scale lexicostatistical language classification will thus become feasible. The ASJP (Automated Similarity Judgment Program) project[50] represents such an approach, where automatically estimated distances between languages are provided as input to phylogenetic programs originally developed in computational biology (Felsenstein 2004), for the purpose of inferring genetic relationships among organisms.

As noted above, traditional lexicostatistics assumes that the cognate judgments for a group of languages have been supplied beforehand. Given a standardized word list, consisting of 40–100 items, the distance between a pair of languages is defined as the percentage of shared cognates subtracted from 100%. This procedure is applied to all pairs of languages under consideration, to produce a pairwise inter-language distance matrix. This inter-language distance matrix is then supplied to a tree-building algorithm such as Neighbor-Joining (NJ; Saitou and Nei 1987) or a clustering algorithm such as Unweighted Pair Group Method with Arithmetic Mean (UPGMA; Sokal and Michener 1958) to infer a tree structure for the set of languages. Swadesh (1950) applies essentially this method – although completely manually – to the Salishan languages. The resulting "family tree" is reproduced in figure 12.1.

The crucial element in these automated approaches is the method used for determining the overall similarity between two word lists.[51] Often, this is some variant of the popular edit distance or Levenshtein distance (LD; Levenshtein 1966). LD for a pair of strings is defined as the minimum number of symbol (character) additions, deletions and substitutions needed to

[50] `http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm`

[51] At this point, we use "word list" and "language" interchangeably. Strictly speaking, a language, as identified by its ISO 639-3 code, can have as many word lists as it has recognized (described) varieties, i.e., *doculects* (Nordhoff and Hammarström 2011).

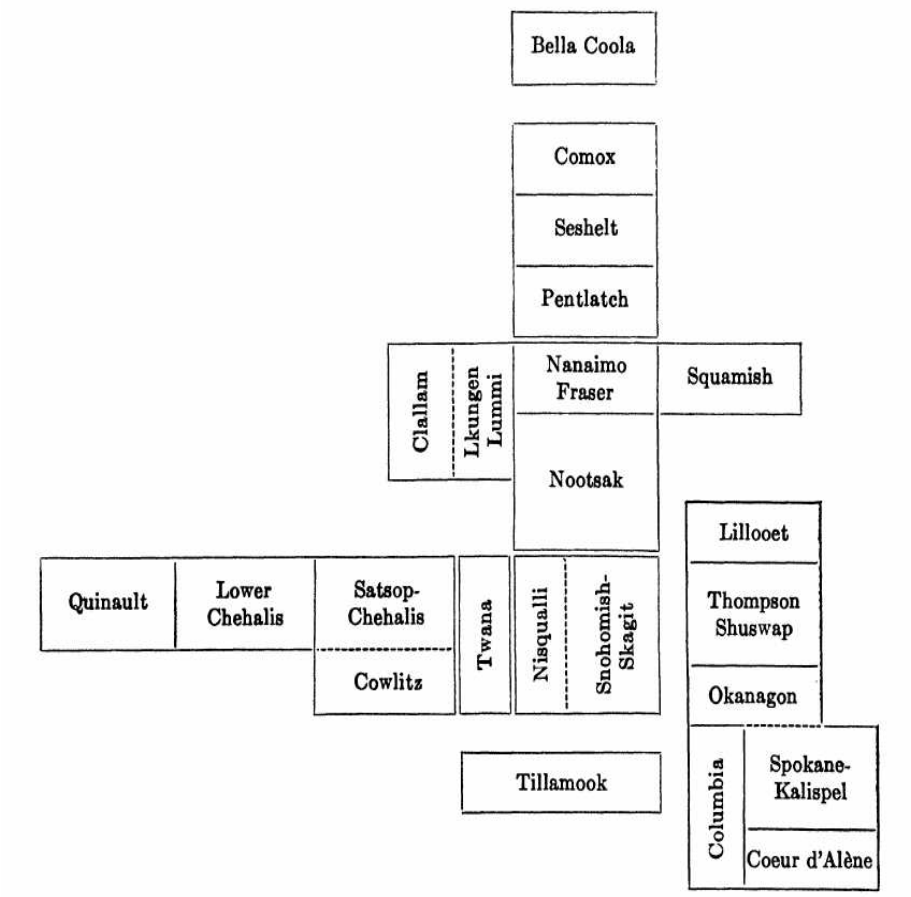


Figure 12.1: Salishan language family box-diagram from Swadesh 1950.

transform one string into the other. A modified LD (called LDND) is used by the ASJP consortium, as reported in their publications (e.g., Bakker et al. 2009 and Holman et al. 2008b).

12.2 Related Work

Cognate identification and tree inference are closely related tasks in historical linguistics. Considering each task as a computational module would mean that each cognate set identified across a set of tentatively related languages feed into the refinement of the tree inferred at each step. In a critical article, Nichols (1996) points out that the historical linguistics enterprise, since its

beginning, always used a refinement procedure to posit relatedness and tree structure for a set of tentatively related languages.[52] The inter-language distance approach to tree-building, is incidentally straightforward and comparably accurate in comparison to the computationally intensive Bayesian-based tree-inference approach of Greenhill and Gray (2009).[53]

The inter-language distances are either an aggregate score of the pairwise item distances or based on a distributional similarity score. The string similarity measures used for the task of cognate identification can also be used for computing the similarity between two lexical items for a particular word sense.

12.2.1 Cognate identification

The task of automatic cognate identification has received a lot of attention in language technology. Kondrak (2002a) compares a number of algorithms based on phonetic and orthographical similarity for judging the cognateness of a word pair. His work surveys string similarity / distance measures such as *edit distance*, *dice coefficient*, and *longest common subsequence ratio* (LCSR) for the task of cognate identification. It has to be noted that, until recently (Hauer and Kondrak 2011; List 2012), most of the work in cognate identification focused on determining the cognateness between a word pair and not among a set of words sharing the same meaning.

Ellison and Kirby (2006) use Scaled Edit Distance (SED)[54] for computing intra-lexical similarity for estimating language distances based on the dataset of Indo-European languages prepared by Dyen, Kruskal and Black (1992). The language distance matrix is then given as input to the NJ algorithm – as implemented in the PHYLIP package (Felsenstein 2002) – to infer a tree for 87 Indo-European languages. They make a qualitative evaluation of the inferred tree against the standard Indo-European tree.

Kondrak (2000) developed a string matching algorithm based on articulatory features (called ALINE) for computing the similarity between a word pair. ALINE was evaluated for the task of cognate identification against machine learning algorithms such as Dynamic Bayesian Networks and

[52] This idea is quite similar to the well-known Expectation-Maximization paradigm in machine learning. Kondrak (2002b) employs this paradigm for extracting sound correspondences by pairwise comparisons of word lists for the task of cognate identification. A recent paper by Bouchard-Côté et al. (2013) employs a feed-back procedure for the reconstruction of Proto-Austronesian with a great success.

[53] For a comparison of these methods, see Wichmann and Rama 2014.

[54] SED is defined as the edit distance normalized by the average of the lengths of the pair of strings.

Pairwise HMMs for automatic cognate identification (Kondrak and Sherif 2006). Even though the approach is technically sound, it suffers due to the very coarse phonetic transcription used in Dyen, Kruskal and Black's Indo-European dataset.[55]

Inkpen, Frunza and Kondrak (2005) compared various string similarity measures for the task of automatic cognate identification for two closely related languages: English and French. The paper shows an impressive array of string similarity measures. However, the results are very language-specific, and it is not clear that they can be generalized even to the rest of the Indo-European family.

Petroni and Serva (2010) use a modified version of Levenshtein distance for inferring the trees of the Indo-European and Austronesian language families. LD is usually normalized by the maximum of the lengths of the two words to account for length bias. The length normalized LD can then be used in computing distances between a pair of word lists in at least two ways: LDN and LDND (Levenshtein Distance Normalized Divided). LDN is computed as the sum of the length normalized Levenshtein distance between the words occupying the same meaning slot divided by the number of word pairs. Similarity between phoneme inventories and chance similarity might cause a pair of not-so related languages to show up as related languages. This is compensated for by computing the length-normalized Levenshtein distance between all the pairs of words occupying different meaning slots and summing the different word-pair distances.

The summed Levenshtein distance between the words occupying the same meaning slots is divided by the sum of Levenshtein distances between different meaning slots. The intuition behind this idea is that if two languages are shown to be similar (small distance) due to accidental chance similarity then the denominator would also be small and the ratio would be high.

If the languages are not related and also share no accidental chance similarity, then the distance as computed in the numerator would be unaffected by the denominator. If the languages are related then the distance as computed in the numerator is small anyway, whereas the denominator would be large since the languages are similar due to genetic relationship and not from chance similarity. Hence, the final ratio would be smaller than the original distance given in the numerator.

Petroni and Serva (2010) claim that LDN is more suitable than LDND for measuring linguistic distances. In reply, Wichmann et al. (2010a) empirically show that LDND performs better than LDN for distinguishing pairs of

[55]The dataset contains 200-word Swadesh lists for 95 language varieties. Available on `http://www.wordgumbo.com/ie/cmp/index.htm`.

languages belonging to the same family from pairs of languages belonging to different families.

As noted by Jäger (2013), Levenshtein distance only matches strings based on symbol identity whereas a graded notion of sound similarity would be a closer approximation to historical linguistics as well as achieving better results at the task of phylogenetic inference. Jäger (2013) uses empirically determined weights between symbol pairs (from computational dialectometry; Wieling, Prokić and Nerbonne 2009) to compute distances between ASJP word lists and finds that there is an improvement over LDND at the task of internal classification of languages.

12.2.2 Distributional similarity measures

Huffman (1998) compute pairwise language distances based on character *n*-grams extracted from Bible texts in European and American Indian languages (mostly from the Mayan language family). Singh and Surana (2007) use character *n*-grams extracted from raw comparable corpora of ten languages from the Indian subcontinent for computing the pairwise language distances between languages belonging to two different language families (Indo-Aryan and Dravidian). Rama and Singh (2009) introduce a factored language model based on articulatory features to induce an articulatory feature level *n*-gram model from the dataset of Singh and Surana 2007. The feature *n*-grams of each language pair are compared using a distributional similarity measure called cross-entropy to yield a single point distance between the language pair. These scholars find that the distributional distances agree with the standard classification to a large extent.

Inspired by the development of tree similarity measures in computational biology, Pompei, Loreto and Tria (2011) evaluate the performance of LDN vs. LDND on the ASJP and Austronesian Basic Vocabulary databases (Greenhill, Blust and Gray 2008). They compute NJ and Minimum Evolution trees[56] for LDN as well as LDND distance matrices. They compare the inferred trees to the classification given in the *Ethnologue* (Lewis 2009) using two different tree similarity measures: Generalized Robinson-Foulds distance (GRF; A generalized version of Robinson-Foulds [RF] distance; Robinson and Foulds 1979) and Generalized Quartet distance (GQD; Christiansen et al. 2006). GRF and GQD are specifically designed to account for the polytomous nature – a node having more than two children – of the Ethnologue trees. For example, the Dravidian family tree shown in figure 12.3 exhibits four branches radiating from the top node. measures Finally, Huff and Lonsdale

[56] A tree building algorithm closely related to NJ.

(2011) compare the NJ trees from ALINE and LDND distance metrics to Ethnologue trees using RF distance. The authors did not find any significant improvement by using a linguistically well-informed similarity measure such as ALINE over LDND.

12.3 Is LD the best string similarity measure for language classification?

LD is only one of a number of string similarity measures used in fields such as language technology, information retrieval, and bio-informatics. Beyond the works cited above, to the best of our knowledge, there has been no study to compare different string similarity measures on something like the ASJP dataset in order to determine their relative suitability for genealogical classification.[57] In this paper we compare various string similarity measures[58] for the task of automatic language classification. We evaluate their effectiveness in language discrimination through a distinctiveness measure; and in genealogical classification by comparing the distance matrices to the language classifications provided by WALS (World Atlas of Language Structures; Haspelmath et al. 2011)[59] and Ethnologue.

Consequently, in this article we attempt to provide answers to the following questions:

- Out of the numerous string similarity measures listed below in section 12.5:
  - Which measure is best suited for the tasks of distinguishing related lanugages from unrelated languages?
  - Which is measure is best suited for the task of internal language classification?
  - Is there a procedure for determining the best string similarity measure?

[57] One reason for this may be that the experiments are computationally demanding, requiring several days for computing a single measure over the whole ASJP dataset.

[58] A longer list of string similarity measures is available on:
`http://www.coli.uni-saarland.de/courses/LT1/2011/slides/stringmetrics.pdf`

[59] WALS does not provide a classification to all the languages of the world. The ASJP consortium gives a WALS-like classification to all the languages present in their database.

12.4 Database and language classifications

12.4.1 Database

The ASJP database offers a readily available, if minimal, basis for massive cross-linguistic investigations. The ASJP effort began with a small dataset of 100-word lists for 245 languages. These languages belong to 69 language families. Since its first version presented by Brown et al. (2008), the ASJP database has been going through a continuous expansion, to include in the version used here (v. 14, released in 2011)[60] more than 5500 word lists representing close to half the languages spoken in the world (Wichmann et al. 2011b). Because of the findings reported by Holman et al. (2008b), the later versions of the database aimed to cover only the 40-item most stable Swadesh sublist, and not the 100-item list.

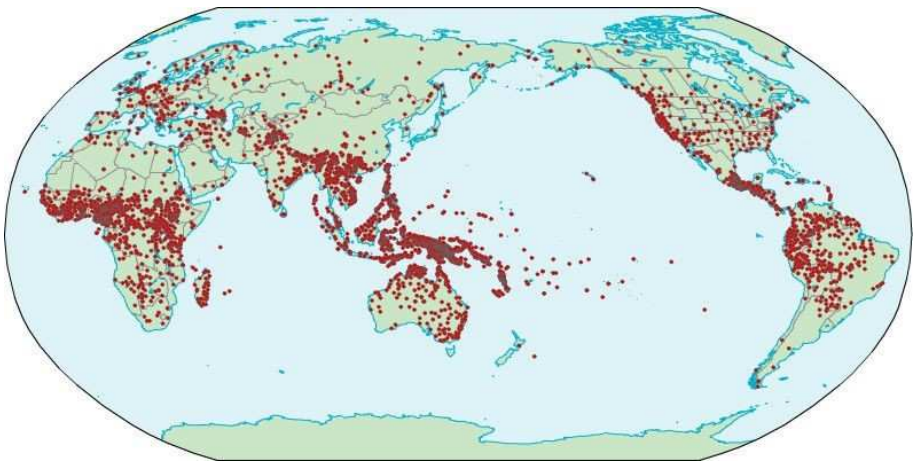


Figure 12.2: Distribution of languages in ASJP database (version 14).

Each lexical item in an ASJP word list is transcribed in a broad phonetic transcription known as ASJP Code (Brown et al. 2008). The ASJP code consists of 34 consonant symbols, 7 vowels, and four modifiers ($*$, ", $\sim$, \$), all rendered by characters available on the English version of the QWERTY keyboard. Tone, stress, and vowel length are ignored in this transcription format. The three modifiers combine symbols to form phonologically complex segments (e.g., aspirated, glottalized, or nasalized segments).

In order to ascertain that our results would be comparable to those published by the ASJP group, we successfully replicated their experiments for LDN and LDND measures using the ASJP program and the ASJP dataset

[60]The latest version is v. 16, released in 2013.

version 12 (Wichmann et al. 2010c).[61] This database comprises reduced (40-item) Swadesh lists for 4169 linguistic varieties. All pidgins, creoles, mixed languages, artificial languages, proto-languages, and languages extinct before 1700 CE were excluded for the experiment, as were language families represented by less than 10 word lists (Wichmann et al. 2010a),[62] as well as word lists containing less than 28 words (70% of 40). This leaves a dataset with 3730 word lists. It turned out that an additional 60 word lists did not have English glosses for the items, which meant that they could not be processed by the program, so these languages were also excluded from the analysis.

All the experiments reported in this paper were performed on a subset of version 14 of the ASJP database whose language distribution is shown in figure 12.2.[63] The database has 5500 word lists. The same selection principles that were used for version 12 (described above) were applied for choosing the languages to be included in our experiments. The final dataset for our experiments has 4743 word lists for 50 language families. We use the family names of the WALS (Haspelmath et al. 2011) classification.

The WALS classification is a two-level classification where each language belongs to a genus and a family. A genus is a genetic classification unit given by Dryer (2000) and consists of set of languages supposedly descended from a common ancestor which is 3000 to 3500 years old. For instance, Indo-Aryan languages are classified as a separate genus from Iranian languages although, it is quite well known that both Indo-Aryan and Iranian languages are descended from a common proto-Indo-Iranian ancestor.

The Ethnologue classification is a multi-level tree classification for a language family. This classification is often criticized for being too "lumping", i.e., too liberal in positing genetic relatedness between languages. The highest node in a family tree is the family itself and languages form the lowest nodes (leaves). A internal node in the tree is not necessarily binary. For instance, the Dravidian language family has four branches emerging from the top node (see figure 12.3 for the Ethnologue family tree of Dravidian languages).

[61] The original python program was created by Hagen Jung. We modified the program to handle the ASJP modifiers.

[62] The reason behind this decision is that correlations resulting from smaller samples (less than 40 language pairs) tend to be unreliable.

[63] Available for downloading at `http://email.eva.mpg.de/~wichmann/listss14.zip`.

Family Name	WN	# WLs	Family Name	WN	# WLs
Afro-Asiatic	AA	287	Mixe-Zoque	MZ	15
Algic	Alg	29	MoreheadU.Maro	MUM	15
Altaic	Alt	84	Na-Dene	NDe	23
Arwakan	Arw	58	Nakh-Daghestanian	NDa	32
Australian	Aus	194	Niger-Congo	NC	834
Austro-Asiatic	AuA	123	Nilo-Saharan	NS	157
Austronesian	An	1008	Otto-Manguean	OM	80
Border	Bor	16	Panoan	Pan	19
Bosavi	Bos	14	Penutian	Pen	21
Carib	Car	29	Quechuan	Que	41
Chibchan	Chi	20	Salish	Sal	28
Dravidian	Dra	31	Sepik	Sep	26
Eskimo-Aleut	EA	10	Sino-Tibetan	ST	205
Hmong-Mien	HM	32	Siouan	Sio	17
Hokan	Hok	25	Sko	Sko	14
Huitotoan	Hui	14	Tai-Kadai	TK	103
Indo-European	IE	269	Toricelli	Tor	27
Kadugli	Kad	11	Totonacan	Tot	14
Khoisan	Kho	17	Trans-NewGuinea	TNG	298
Kiwain	Kiw	14	Tucanoan	Tuc	32
LakesPlain	LP	26	Tupian	Tup	47
Lower-Sepik-Ramu	LSR	20	Uralic	Ura	29
Macro-Ge	MGe	24	Uto-Aztecan	UA	103
Marind	Mar	30	West-Papuan	WP	33
Mayan	May	107	WesternFly	WF	38

Table 12.1: Distribution of language families in ASJP database. WN and WLs stands for WALS Name and Word Lists.

12.5 Similarity measures

For the experiments decribed below, we have considered both string similarity measures and distributional measures for computing the distance between a pair of languages. As mentioned earlier, string similarity measures work at the level of word pairs and provide an aggregate score of the similarity between word pairs whereas distributional measures compare the n-gram profiles between a language pair to yield a distance score.

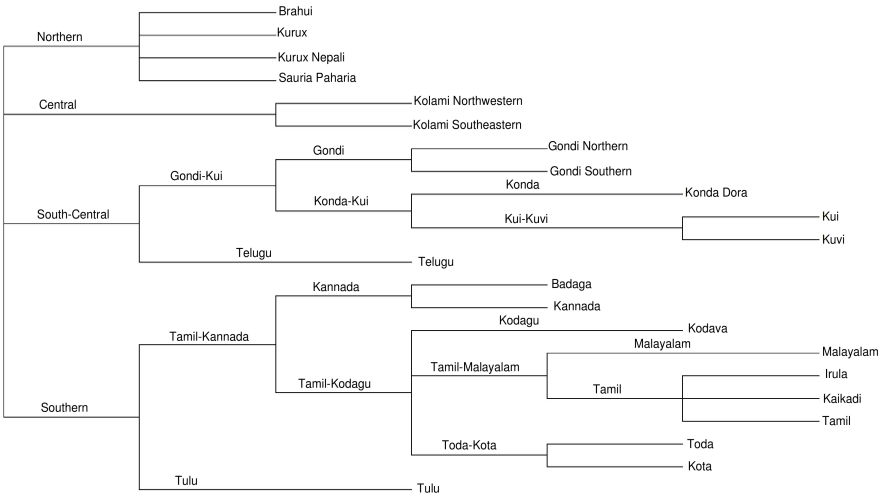


Figure 12.3: Ethnologue tree for the Dravidian language family.

12.5.1 String similarity measures

The different string similarity measures for a word pair that we have investuigated are the following:

- *IDENT* returns 1 if the words are identical, otherwise it returns 0.
- *PREFIX* returns the length of the longest common prefix divided by the length of the longer word.
- *DICE* is defined as the number of shared bigrams divided by the total number of bigrams in both the words.
- *LCS* is defined as the length of the longest common subsequence divided by the length of the longer word (Melamed 1999).
- *TRIGRAM* is defined in the same way as *DICE* but uses trigrams for computing the similarity between a word pair.
- *XDICE* is defined in the same way as *DICE* but uses "extended bigrams", which are trigrams without the middle letter (Brew and McKelvie 1996).
- Jaccard's index, *JCD*, is a set cardinality measure that is defined as the ratio of the number of shared bigrams between the two words to the ratio of the size of the union of the bigrams between the two words.

- *LDN*, as defined above.

Each word-pair similarity score is converted to its distance counterpart by subtracting the score from 1.0.[64] Note that this conversion can sometimes result in a negative distance which is due to the double normalization involved in LDND.[65] This distance score for a word pair is then used to compute the pairwise distance between a language pair. The distance computation between a language pair is performed as described in section 12.2.1. Following the naming convention of LDND, a suffix "D" is added to the name of each measure to indicate its LDND distance variant.

12.5.2 N-gram similarity

N-gram similarity measures are inspired by a line of work initially pursued in the context of information retrieval, aiming at automatic language identification in a multilingual document. Cavnar and Trenkle (1994) used character *n*-grams for text categorization. They observed that different document categories – including documents in different languages – have characteristic character *n*-gram profiles. The rank of a character *n*-gram varies across different categories and documents belonging to the same category have similar character *n*-gram Zipfian distributions.

Building on this idea, Dunning (1994, 1998) postulates that each language has its own signature character (or phoneme; depending on the level of transcription) *n*-gram distribution. Comparing the character *n*-gram profiles of two languages can yield a single point distance between the language pair. The comparison procedure is usually accomplished through the use of one of the distance measures given in Singh 2006. The following steps are followed for extracting the phoneme *n*-gram profile for a language:

- An *n*-gram is defined as the consecutive phonemes in a window of *N*. The value of *N* usually ranges from 1 to 5.
- All *n*-grams are extracted for a lexical item. This step is repeated for all the lexical items in a word list.
- All the extracted *n*-grams are mixed and sorted in the descending order of their frequency. The relative frequency of the *n*-grams are computed.

[64] Lin (1998) investigates three distance to similarity conversion techniques and motivates the results from an information-theoretical point of view. In this article, we do not investigate the effects of similarity to distance conversion. Rather, we stick to the traditional conversion technique.

[65] Thus, the resulting distance is not a true distance metric.

- Only the top *G* *n*-grams are retained and the rest of them are discarded. The value of *G* is determined empirically.

For a language pair, the *n*-gram profiles can be compared using one of the following distance measures:

1. *Out-of-Rank measure* is defined as the aggregate sum of the absolute difference in the rank of the shared *n*-grams between a pair of languages. If there are no shared bigrams between an *n*-gram profile, then the difference in ranks is assigned a maximum out-of-place score.

2. *Jaccard's index* is a set cardinality measure. It is defined as the ratio of the cardinality of the intersection of the *n*-grams between the two languages to the cardinality of the union of the two languages.

3. *Dice distance* is related to Jaccard's Index. It is defined as the ratio of twice the number of shared *n*-grams to the total number of *n*-grams in both the language profiles.

4. *Manhattan distance* is defined as the sum of the absolute difference between the relative frequency of the shared *n*-grams.

5. *Euclidean distance* is defined in a similar fashion to Manhattan distance where the individual terms are squared.

While replicating the original ASJP experiments on the version 12 ASJP database, we also tested if the above distributional measures, [1–5] perform as well as LDN. Unfortunately, the results were not encouraging, and we did not repeat the experiments on version 14 of the database. One main reason for this result is the relatively small size of the ASJP concept list, which provides a poor estimate of the true language signatures.

This factor speaks equally, or even more, against including another class of *n*-gram-based measures, namely information-theoretic measures such as *cross entropy* and *KL-divergence*. These measures have been well-studied in natural language processing tasks such as machine translation, natural language parsing, sentiment identification, and also in automatic language identification. However, the probability distributions required for using these measures are usually estimated through maximum likelihood estimation which require a fairly large amount of data, and the short ASJP concept lists will hardly qualify in this regard.

12.6 Evaluation measures

The measures which we have used for evaluating the performance of string similarity measures given in section 12.5 are the following three:

1. *dist* was originally suggested by Wichmann et al. (2010a), and tests if LDND is better than LDN at the task of distinguishing related languages from unrelated languages.

2. *RW* is a special case of Pearson's **r** – called point biserial correlation (Tate 1954) – computes the agreement between a the intra-family pairwise distances and the WALS classification for the family.

3. γ is related to Goodman and Kruskal's Gamma (1954) and measures the strength of association between two ordinal variables. In this paper, it is used to compute the level of agreement between the pairwise intra-family distances and the family's Ethnologue classification.

12.6.1 Distinctiveness measure (dist)

The *dist* measure for a family consists of three components: the mean of the pairwise distances inside a language family (d_{in}); and the mean of the pairwise distances from each language in a family to the rest of the language families (d_{out}). sd_{out} is defined as the standard deviation of all the pairwise distances used to compute d_{out}. Finally, *dist* is defined as $\frac{d_{in}-d_{out}}{sd_{out}}$. The resistance of a string similarity measure to other language families is reflected by the value of sd_{out}.

A comparatively higher *dist* value suggests that a string similarity measure is particularly resistant to random similarities between unrelated languages and performs well at distinguishing languages belonging to the same language family from languages in other language families.

12.6.2 Correlation with WALS

The WALS database provides a three-level classification. The top level is the language family, second level is the genus and the lowest level is the language itself. If two languages belong to different families, then the distance is 3. Two languages that belong to different genera in the same family have a distance of 2. If the two languages fall in the same genus, they have a distance of 1. This allows us to define a distance matrix for each family based on WALS.

The WALS distance matrix can be compared to the distance matrices of any string similarity measure using point biserial correlation – a special case of Pearson's **r**. If a family has a single genus in the WALS classification there is no computation of RW and the corresponding row for a family is empty in table C.3.

12.6.3 Agreement with Ethnologue

Given a distance-matrix d of order $N \times N$, where each cell d_{ij} is the distance between two languages i and j; and an Ethnologue tree E, the computation of γ for a language family is defined as follows:

1. Enumerate all the triplets for a language family of size N. A triplet, t for a language family is defined as $\{i, j, k\}$, where $i \neq j \neq k$ are languages belonging to a family. A language family of size N has $\binom{N}{3}$ triplets.

2. For the members of each such triplet t, there are three lexical distances d_{ij}, d_{ik}, and d_{jk}. The expert classification tree E can treat the three languages $\{i, j, k\}$ in four possible ways (| denotes a partition): $\{i, j \mid k\}$, $\{i, k \mid j\}$, $\{j, k \mid i\}$ or can have a tie where all languages emanate from the same node. All ties are ignored in the computation of γ.[66]

3. A distance triplet d_{ij}, d_{ik}, and d_{jk} is said to agree completely with an Ethnologue partition $\{i, j \mid k\}$ when the following conditions are satisfied:

$$d_{ij} < d_{ik} \tag{12}$$

$$d_{ij} < d_{jk} \tag{13}$$

   A triplet that satisfies these conditions is counted as a concordant comparison, C; else it is counted as a discordant comparison, D.

4. Steps 2 and 3 are repeated for all the $\binom{N}{3}$ triplets to yield γ for a family defined as $\gamma = \frac{C-D}{C+D}$. γ lies in the range $[-1, 1]$ where a score of -1 indicates perfect disagreement and a score of $+1$ indicates perfect agreement.

[66] We do not know what a tie in the gold standard indicates: uncertainty in the classification, or a genuine multi-way branching? Whenever the Ethnologue tree of a family is completely unresolved, it is shown by an empty row. For example, the family tree of Bosavi languages is a star structure. Hence, the corresponding row in table C.1 is left empty.

At this point, one might wonder about the decision for not using an off-the-shelf tree-building algorithm to infer a tree and compare the resulting tree with the Ethnologue classification. Although both Pompei, Loreto and Tria (2011) and Huff and Lonsdale (2011) compare their inferred trees – based on Neighbor-Joining and Minimum Evolution algorithms – to Ethnologue trees using cleverly crafted tree-distance measures (GRF and GQD), they do not make the more intuitively useful direct comparison of the distance matrices to the Ethnologue trees. The tree inference algorithms use heuristics to find the best tree from the available tree space. The number of possible rooted, non-binary and unlabeled trees is quite large even for a language family of size 20 – about 256×10^6.

A tree inference algorithm uses heuristics to reduce the tree space to find the best tree that explains the distance matrix. A tree inference algorithm can make mistakes while searching for the best tree. Moreover, there are many variations of Neighbor-Joining and Minimum Evolution algorithms.[67] Ideally, one would have to test the different tree inference algorithms and then decide the best one for our task. However, the focus of this paper rests on the comparison of different string similarity algorithms and not on tree inference algorithms. Hence, a direct comparison of a family's distance matrix to the family's Ethnologue tree circumvents the choice of the tree inference algorithm.

12.7 Results and discussion

In table 12.2 we give the results of our experiments. We only report the average results for all measures across the families listed in table 12.1. Further, we check the correlation between the performance of the different string similarity measures across the three evaluation measures by computing Spearman's ρ. The pairwise ρ is given in table 12.3. The high correlation value of 0.95 between RW and γ suggests that all the measures agree roughly on the task of internal classification.

[67] `http://www.atgc-montpellier.fr/fastme/usersguide.php`

Measure	Average Dist	Average RW	Average γ
DICE	3.3536	0.5449	0.6575
DICED	9.4416	0.5495	0.6607
IDENT	1.5851	0.4013	0.2345
IDENTD	8.163	0.4066	0.3082
JCD	13.9673	0.5322	0.655
JCDD	15.0501	0.5302	0.6622
LCS	3.4305	0.6069	0.6895
LCSD	6.7042	0.6151	0.6984
LDN	3.7943	0.6126	0.6984
LDND	7.3189	0.619	0.7068
PREFIX	3.5583	0.5784	0.6747
PREFIXD	7.5359	0.5859	0.6792
TRIGRAM	1.9888	0.4393	0.4161
TRIGRAMD	9.448	0.4495	0.5247
XDICE	0.4846	0.3085	0.433
XDICED	2.1547	0.4026	0.4838
Average	6.1237	0.5114	0.5739

Table 12.2: Average results for each string similarity measure across the 50 families. The rows are sorted by the name of the measure.

	Dist	RW
γ	0.30	0.95
Dist		0.32

Table 12.3: Spearman's ρ between γ, RW and Dist

The average scores in each column suggest that the string similarity measures exhibit different degrees of performance. How does one decide which measure is the best in a column? What kind of statistical testing procedure should be adopted for deciding upon a measure? We address this questions through the following procedure:

1. For a column i, sort the average scores, s in descending order.
2. For a row index $1 \leq r \leq 16$, test the significance of $s_r \geq s_{r+1}$ through a sign test (Sheskin 2003). This test yields a $p-value$.

The above significant tests are not independent by themselves. Hence, we cannot reject a null hypothesis H_0 at a significance level of $\alpha = 0.01$. The α

needs to be corrected for multiple tests. Unfortunately, the standard Bonferroni's multiple test correction or Fisher's Omnibus test works for a global null hypothesis and not at the level of a single test. We follow the procedure, called False Discovery Rate (FDR), given by Benjamini and Hochberg (1995) for adjusting the α value for multiple tests. Given $H_1 \ldots H_m$ null hypotheses and $P_1 \ldots P_m$ p-values, the procedure works as follows:

1. Sort the P_k, $1 \leq k \leq m$, values in ascending order. k is the rank of a p-value.
2. The adjusted α_k^* value for P_k is $\frac{k}{m}\alpha$.
3. Reject all the H_0s from $1, \ldots, k$ where $P_{k+1} > \alpha_k^*$.

The above procedure ensures that the chance of incorrectly rejecting a null hypothesis is 1 in 20 for $\alpha = 0.05$ and 1 in 100 for $\alpha = 0.01$. In this experimental context, this suggests that we erroneously reject 0.75 true null hypotheses out of 15 hypotheses for $\alpha = 0.05$ and 0.15 hypotheses for $\alpha = 0.01$. We report the Dist, γ, and RW for each family in tables C.1, C.2, and C.3. In each of these tables, only those measures which are above the average scores from table 12.2, are reported.

The FDR procedure for γ suggests that no sign test is significant. This is in agreement with the result of Wichmann et al. 2010a, who showed that the choice of LDN or LDND is quite unimportant for the task of internal classification. The FDR procedure for RW suggests that LDN $>$ LCS, LCS $>$ PREFIXD, DICE $>$ JCD, and JCD $>$ JCDD. Here $A > B$ denotes that A is significantly better than B. The FDR procedure for Dist suggests that JCDD $>$ JCD, JCD $>$ TRID, DICED $>$ IDENTD, LDND $>$ LCSD, and LCSD $>$ LDN.

The results point towards an important direction in the task of building computational systems for automatic language classification. The pipeline for such a system consists of (1) distinguishing related languages from unrelated languages; and (2) internal classification accuracy. JCDD performs the best with respect to Dist. Further, JCDD is derived from JCD and can be computed in $\mathcal{O}(m+n)$, for two strings of length m and n. In comparison, LDN is in the order of $\mathcal{O}(mn)$. In general, the computational complexity for computing distance between two word lists for all the significant measures is given in table 12.4. Based on the computational complexity and the significance scores, we propose that JCDD be used for step 1 and a measure like LDN be used for internal classification.

Measure	Complexity
JCDD	$C\mathcal{O}(m+n+\min(m-1,n-1))$
JCD	$l\mathcal{O}(m+n+\min(m-1,n-1))$
LDND	$C\mathcal{O}(mn)$
LDN	$l\mathcal{O}(mn)$
PREFIXD	$C\mathcal{O}(\max(m,n))$
LCSD	$C\mathcal{O}(mn)$
LCS	$l\mathcal{O}(mn)$
DICED	$C\mathcal{O}(m+n+\min(m-2,n-2))$
DICE	$l\mathcal{O}(m+n+\min(m-2,n-2))$

Table 12.4: Computational complexity of top performing measures for computing distance between two word lists. Given two word lists each of length l. m and n denote the lengths of a word pair w_a and w_b and $C = l(l-1)/2$

12.8 Conclusion

In this article, we have presented the first known attempt to apply more than 20 different similarity (or distance) measures to the problem of genetic classification of languages on the basis of Swadesh-style core vocabulary lists. The experiments were performed on the wide-coverage ASJP database (about half the world's languages).

We have examined the various measures at two levels, namely: (1) their capability of distinguishing related and unrelated languages; and (2) their performance as measures for internal classification of related languages. We find that the choice of string similarity measure (among the tested pool of measures) is not very important for the task of internal classification whereas the choice affects the results of discriminating related languages from unrelated ones.

13 COGNATE IDENTIFICATION WITH GAP-WEIGHTED STRING SUBSEQUENCES

Rama, Taraka 2015. Automatic cognate identification with gap-weighted string subsequences. *Proceedings of North American Association for Computational Linguistics.* 1227–1231.

Abstract

In this paper, we describe the problem of cognate identification in NLP. We introduce the idea of gap-weighted subsequences for discriminating cognates from non-cognates. We also propose a scheme to integrate phonetic features into the feature vectors for cognate identification. We show that subsequence based features perform better than state-of-the-art classifier for the purpose of cognate identification. The contribution of this paper is the use of subsequence features for cognate identification.

13.1 Introduction

Cognates are words across languages whose origin can be traced back to a common ancestral language. For example, English $\sim$ German *night* $\sim$ *Nacht* 'night' and English *hound* $\sim$ German *Hund* 'dog' are cognates whose origin can be traced back to Proto-Germanic. Sometimes, cognates are not revealingly similar but have changed substantially over time such that they do not share form similarity. An example of such a cognate pair is the English *wheel* and Sanskrit *chakra* 'wheel', which can be traced back to Proto-Indo-European (PIE) $*k^wek^welo$.

Automatic cognate identification, in NLP, refers to the application of string similarity or phonetic similarity algorithms either independently, or in tandem with machine learning algorithms for determining if a given word pair

is cognate or not (Inkpen, Frunza and Kondrak 2005). In NLP, even borrowed words (*loanwords*) are referred to as cognates. In contrast, historical linguistics makes a stark distinction between loanwords and cognates. For example, English *beef* is a loanword from Norman French.

In this paper, we use cognates to refer to those words whose origin can be traced back to a common ancestor. We use string subsequence based features (motivated from string kernels) for automatic cognate identification. We show that subsequence-based features outperform word similarity measures at the task of automatic cognate identification. We motivate the use of subsequence based features in terms of linguistic examples and then proceed to formulate the subsequence based features that can be derived from string kernels (Shawe-Taylor and Cristianini 2004). In information retrieval literature, string subsequences go under the name of skip-grams (Järvelin, Järvelin and Järvelin 2007).

13.2 Related work

The approaches developed by Kondrak and Sherif (2006) and Inkpen, Frunza and Kondrak (2005) supply different string distances between a pair of words as features to a linear classifier. Usually, a linear classifier such as SVM is trained on labeled positive ("cognates") and negative ("non-cognates") examples and tested on a held-out dataset. Basic vocabulary lists such as the ones devised by Morris Swadesh (Swadesh 1952), provide a suitable testing ground for applying machine learning algorithms to automatically identify cognates. Some standardized word lists come with cognate information and, subsequently, can used to infer the relationship between the languages (Dyen, Kruskal and Black 1992).

Ellison and Kirby (2006) use scaled edit distance (normalized by average length) to measure the intra-lexical divergence in a language. The inter-language distance matrix is supplied to a clustering algorithm to infer a tree for the Indo-European language family. The authors only perform a qualitative evaluation of the inferred tree. The authors mention string kernels but do not pursue this line of research further.

Bouchard-Côté et al. (2013) employ a graphical model to reconstruct the proto-word forms from the synchronic word-forms for the Austronesian language family. They compare their automated reconstructions with the ones reconstructed by historical linguists and find that their model beats an edit-distance baseline. However, their model has a requirement that the tree structure between the languages under study has to be known beforehand.

Hauer and Kondrak (2011) – referred to as HK – supply different string similarity scores as features to a SVM classifier for determining if a given word pair is a cognate or not. The authors also employ an additional binary language-pair feature – that is used to weigh the language distance – and find that the additional feature assists the task of semantic clustering of cognates. In this task, the cognacy judgments given by a linear classifier is used to flat cluster the lexical items belonging to a single concept. The clustering quality is evaluated against the gold standard cognacy judgments. Unfortunately, the experiments of these scholars cannot be replicated since the partitioning details of their training and test datasets is not available.

In our experiments, we compare our system's performance with the performance of the classifiers trained from HK-based features. In the next section, we will describe string similarity measures, subsequences features, dataset, and results.

13.3 Cognate identification

13.3.1 String similarity features and issues

Edit distance counts the minimum number of insertions, deletions, and substitutions required to transform one word into another word. Identical words have 0 edit distance. For example, the edit distance between two cognates English *hound* and German *hund* is 1. Similarly, the edit distance between Swedish *i* and Russian *v* 'in', which are cognates, is 1. The edit distance treats both of the cognates at the same level and does not reflect the amount of change which has occurred in the Swedish and Russian words from the PIE word.

Dice is another string similarity measure that defines similarity between two strings as the ratio between the number of common bigrams to the total number of bigrams. The similarity between Lusatian *dolhi* and Czech *dluhe* 'long' is 0 since they do not share any common bigrams and the edit distance between the two strings is 3. Although the two words share all the consonants, the Dice score is 0 due to the intervening vowels.

Another string similarity measure, Longest Common Subsequence (LCS) measures the length of the longest common subsequence between the two words. The LCS is 4 (*hund*), 0 (*i* and *v*), and 3 (*dlh*) for the above examples. One can put forth a number of examples which are problematical for the commonly-used string similarity measures. Alternatively, string kernels in machine learning research offer a way to exploit the similarities between two words without any restrictions on the length and character similarity.

13.3.2 Subsequence features

Subsequences as formulated below weigh the similarity between two words based on the number of dropped characters and combine phoneme classes seamlessly. Having motivated why subsequences seems to be a good idea, we formulate subsequences below.

We follow the notation given in Shawe-Taylor and Cristianini 2004 to formulate our representation of a string (word). Let Σ denote the set of phonetic alphabet. Given a string s over Σ, the subsequence vector $\Phi(s)$ is defined as follows. The string s can be decomposed as $s_1,\ldots,s_{|s|}$ where $|s|$ denotes the length of the string. Let $\overrightarrow{I}$ denote a sequence of indices $(i_1,\ldots,i_{|u|})$ where, $1 \leq i_1 < \ldots < i_{|u|} \leq |s|$. Then, a subsequence u is a sequence of characters $s[\overrightarrow{I}]$. Note that a subsequence can occur multiple times in a string. Then, the weight of u, $\phi_u(s)$ is defined as $\sum_{\overrightarrow{I}:u=s[\overrightarrow{I}]} \lambda^{l(\overrightarrow{I})}$ where, $l(\overrightarrow{I}) = i_{|u|} - i_1 + 1$ and $\lambda \in (0,1)$ is a decay factor.

The subsequence vector $\Phi(s)$ is composed of $\phi_u(s)\ \forall u \in \bigcup_{n=1}^{p} \Sigma^n$, where $1 \leq n \leq p$ is the length of u and p is the maximum length of the subsequences. As $\lambda \to 0$, a subsequence is constrained to a substring. As $\lambda \to 1$, $\phi_u(s)$ counts the frequency of u in s. We also experiment with different values of λ in this paper.

The λ factor is exponential and penalizes u over long gaps in a string. Due to the above formulation, the frequency of a subsequence u in a single string is also taken into account. The subsequence formulation also allows for the incorporation of a class-based features easily. For instance, each σ in u can be mapped to its **C**onsonant/**V**owel class: $\sigma \mapsto \{C,V\}$. The subsequence formulation also allows us to map each phonetic symbol (for example, from International Phonetic Alphabet [IPA]) to an intermediary phonetic alphabet also. Unfortunately, the current dataset is not transcribed in IPA to convert it into an intermediary broader format. In this paper, we map each string s into its C,V sequence s_{cv} and then compute the subsequence weights.[68] A combined subsequence vector $\Phi(s+s_{cv})$ is further normalized by its norm, $|\Phi(s+s_{cv})|$, to transform into a unit vector. The common subsequence vector $\Phi(s_1,s_2)$ is composed of all the common subsequences between s_1 and s_2. The weight of a common subsequence is $\phi_u(s_1)+\phi_u(s_2)$.

Moschitti, Ju and Johansson (2012) list the features of the above weighting scheme.

- Longer subsequences receive lower weights.
- Characters can be omitted (called *gaps*).
- The exponent of λ penalizes recurring subsequences with longer gaps.

[68] $V = \{a,e,i,o,u,y\}$, $C = \Sigma \setminus V$.

For a string of length m and subsequence length n, the computational complexity is in the order of $\mathcal{O}(mn)$.

On a linguistic note, gaps are consistent with the prevalent sound changes such as sound loss, sound gain, and assimilation,[69] processes which alter word forms in an ancestral language causing the daughter languages to have different surface forms. The λ factor weighs the number of gaps found in a subsequence. For instance, the Sardinian word form for 'fish' *pissi* has the subsequence *ps* occurring twice but with different weights: λ^3, λ^4. Hence, *ps*'s weight is $\lambda^3 + \lambda^4$. On another note, the idea of gap subsequences subsumes the definitions of different n-gram similarities introduced by Kondrak (2005).

The combined feature vector, for a word pair, is used to train a SVM classifier. In our experiments, we use the LIBLINEAR package (Fan et al. 2008) to solve the primal problem with L_2-regularization and L_2-loss. The next subsection describes the makeup of the dataset. We use the default SVM parameters since we did not observe any difference in our development experiments.

13.3.3 Dataset and results

In this section, we will present the dataset, HK features, and results of our experiments.

13.3.3.1 Dataset.

We used the publicly available[70] Indo-European dataset (Dyen, Kruskal and Black 1992) for our experiments. The dataset has 16,520 lexical items for 200 concepts and 84 language varieties. Each word form is assigned to a unique CCN (Cognate Class Number). There are more than 200 identical non-cognate pairs in the dataset. For the first experiment, we extracted all word pairs for a concept and assigned a positive label if the word pair has an identical CCN; a negative label, if the word pair has different CCNs. We extracted a total of 619,635 word pairs out of which 145,145 are cognates. The dataset is transcribed in a broad romanized phonetic alphabet.

We explored if we could use two other word list databases: ASJP (Brown et al. 2008) and Ringe, Warnow and Taylor (2002) for our experiments. Although the ASJP database has word lists for more than half of the world's

[69] A sound can assimilate to a neighboring sound. Sanskrit *agni* > Prakrit *aggi* 'fire'. Compare the Latin form *ignis* with the Sanskrit form.

[70] http://www.wordgumbo.com/ie/cmp/iedata.txt

languages, it has cognacy judgments for few selected languages and is limited to 40 concepts. Moreover, the ASJP database does not have cognacy judgments for Indo-European family. The other dataset of Ringe, Warnow and Taylor (2002) has items for 24 Indo-European languages which are transcribed in an orthographic format and not in a uniform phonetic alphabet.[71] Moreover, there are a number of missing items for some of the languages. Hence, we did not use Ringe et al.'s dataset in our experiments. In contrast, Dyen's dataset is much larger and transcribed in an uniform format. Now, we proceed to describe the previous best-performing system.

13.3.3.2 HK's system.

We compare the performance of subsequence features against the SVM classifier system trained on the following word-similarity features from Hauer and Kondrak 2011:

- Edit distance.
- Length of longest common prefix.
- Number of common bigrams.
- Lengths of individual words.
- Absolute difference between the lengths of the words.

13.3.3.3 Cross-Validation experiment.

As a first step, we perform a random ten-fold cross-validation of the dataset and report the accuracies for various values of λ and p. The results of this experiment are shown in figure 13.1. The best results are obtained at $\lambda = 0.8, p = 3$. The accuracies increase with an increment in the value of λ until 0.8 for all $p > 1$ (non-unigram models). This experiment is mainly designed to test the robustness of subsequence features against random splits in the dataset which turns out to be robust. The subsequence features outperform HK-based classifier in this experiment.

	positive	negative
training	111,918	353,957
test	33,227	120,533

Table 13.1: Number of positive and negative examples in the training and test sets. The ratio of positive to negative examples is 1 : 3.62.

[71] http://www.cs.rice.edu/ nakhleh/CPHL/ie-wordlist-07.pdf

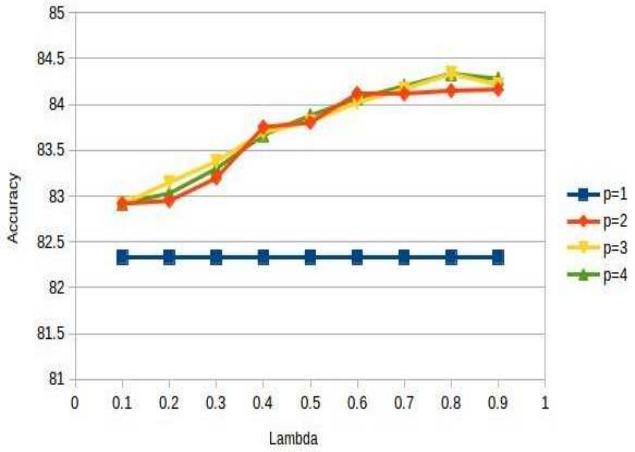


Figure 13.1: Ten-fold cross-validation accuracy for incremental λ and p. The accuracy of the system of HK is 82.61%.

13.3.3.3.1 Concepts experiment.

In this experiment, we split our dataset into two sets by concepts; and train on one set and test on the other. To replicate our dataset, we performed an alphabetical sort of the concepts and split the concepts into training and testing datasets with a ratio of 3 : 1. Now, we extract positive and negative examples from each subset of concepts; and train and test on each concepts' subset. We also performed a 3-fold cross-validation on the training set to tune c (SVM hyperparameter). We observed that the value of c did not effect the cross-validation accuracy on the training dataset. Hence we fixed c at 1. We also experimented with radial-basis function kernel and polynomial kernels but did not find any improvement over the linear kernel classifier. The composition of the training and test sets is given in table 13.1.

In this experiment, we report the F_1-score, defined as $\frac{2PR}{P+R}$ (**P**recision and **R**ecall), for different values of λ and p. The results of this experiment are shown in figure 13.2. The F_1-score of the system of HK is 0.46 whereas the best performing subsequence system ($\lambda = 0.7, p = 2$) has a score of 0.5. Our system performs better than the system of HK in terms of cross-validation accuracy as well as F_1-score. Overall, all non-unigram models perform better than the system of HK at cross-validation and concepts experiments.

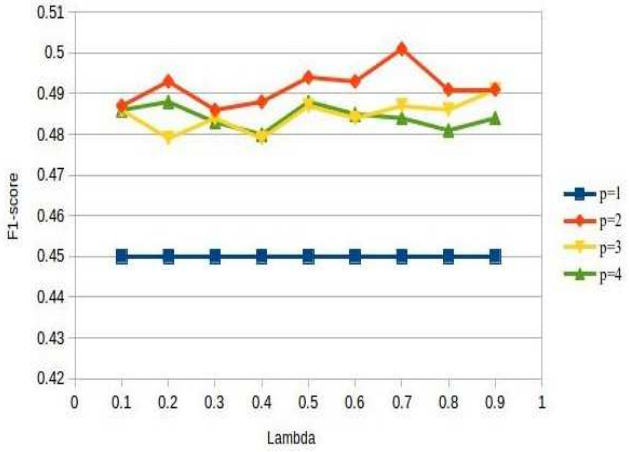


Figure 13.2: F_1-score for different values of p and λ. The F_1-score of the system of HK is 0.46.

13.4 Conclusion

In this paper, we proposed a string kernel based approach for the purpose of cognate identification. We formulated an approach to integrate phonetic features of a phonetic symbol into the feature vector and showed that it beats the system of HK at cognate identification at cross-validation and concepts subsets experiments.

In future, we plan to make a larger dataset of cognacy judgments for other language families in a richer phonetic transcription and integrate articulatory phonetic features into the feature vectors for the purpose of cognate identification. We also plan on testing with different feature vector combinations.

REFERENCES

Abney, Steven 2004. Understanding the Yarowsky algorithm. *Computational Linguistics* 30 (3): 365–395.

Abney, Steven 2010. *Semisupervised learning for computational linguistics*. Chapman & Hall/CRC.

Abney, Steven and Steven Bird 2010. The human language project: Building a universal corpus of the world's languages. *Proceedings of the 48th meeting of the ACL*, 88–97. Uppsala: ACL.

Adesam, Yvonne, Malin Ahlberg and Gerlof Bouma 2012. bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa... Towards lexical link-up for a corpus of Old Swedish. *Proceedings of KONVENS*, 365–369.

Agarwal, Abhaya and Jason Adams 2007. Cognate identification and phylogenetic inference: Search for a better past. Technical Report, Carnegie Mellon University.

Amorim, Carlos Eduardo Guerra, Rafael Bisso-Machado, Virginia Ramallo, Maria Cátira Bortolini, Sandro Luis Bonatto, Francisco Mauro Salzano and Tábita Hünemeier 2013. A Bayesian approach to genome/linguistic relationships in Native South Americans. *PloS one* 8 (5): e64099.

Anttila, Raimo 1989. *Historical and comparative linguistics*. Volume 6 of *Current Issues in Linguistic Theory*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Asher, Ronald E. 2008. Language in historical context. Braj B. Kachru, Yamuna Kachru and S. N. Sridhar (eds), *Language in South Asia*, 31–46. Cambridge: Cambridge University Press.

Atkinson, Quentin, Geoff Nicholls, David Welch and Russell Gray 2005. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103 (2): 193–219.

Atkinson, Quentin D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332 (6027): 346.

Atkinson, Quentin D. and Russell D. Gray 2005. Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic Biology* 54 (4): 513–526.

Atkinson, Quentin D. and Russell D. Gray 2006. How old is the Indo-European language family? Progress or more moths to the flame. Peter Forster and Collin Renfrew (eds), *Phylogenetic methods and the prehistory of languages*, 91–109. Cambridge: The McDonald Institute for Archaelogical Research.

Baayen, R Harald 2009. languager: Data sets and functions with "analyzing linguistic data: A practical introduction to statistics". *R package version*, vol. 1.

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant and Eric W. Holman 2009. Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology* 13 (1): 169–181.

Balanovsky, Oleg, Olga Utevska and Elena Balanovska 2013. Genetics of Indo-European populations: the past, the future. *Journal of Language Relationship* 9: 23–35.

Bates, Douglas and Martin Maechler 2009. lme4: Linear mixed-effects models using S4 classes.

Bauer, Laurie 2007. *The linguistics student's handbook*. Edinburgh: Edinburgh University Press.

Benjamini, Yoav and Yosef Hochberg 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

Bergsland, Knut and Hans Vogt 1962. On the validity of glottochronology. *Current Anthropology* 3 (2): 115–153.

Bickel, Balthasar 2002. The AUTOTYP research program. Invited talk given at the Annual Meeting of the Linguistic Typology Resource Center Utrecht.

Bickel, Balthasar and Johanna Nichols 2002. Autotypologizing databases and their use in fieldwork. *Proceedings of the LREC 2002 workshop on resources and tools in field linguistics*.

Birch, Alexandra, Miles Osborne and Philipp Koehn 2008. Predicting success in machine translation. *Proceedings of the 2008 conference on empirical methods in natural language processing*, 745–754. Honolulu, Hawaii: Association for Computational Linguistics.

Blench, Roger 2013. Why is Africa so linguistically undiverse? exploring substrates and isolates. *Journal of the Association for the Study of Language in Prehistory* XVIII: 43–78.

Bloch, Jules 1954. *The grammatical structure of Dravidian languages*. Pune: Deccan College. Authorized translation from the original French by Ramkrishan Ganesh Harshé.

Bloomfield, Leonard 1935. *Language*. London: Allen, George and Unwin.

Blust, Robert 2011. Austronesian: A sleeping giant? *Language and Linguistics Compass* 5 (8): 538–550.

Bolhuis, Johan J, Ian Tattersall, Noam Chomsky and Robert C Berwick 2014. How could language have evolved? *PLoS biology* 12 (8): e1001934.

Borin, Lars 1988. A computer model of sound change: An example from Old Church Slavic. *Literary and Linguistic Computing* 3 (2): 105–108.

Borin, Lars 2009. Linguistic diversity in the information society. *Proceedings of the SALTMIL 2009 workshop on information retrieval and information extraction for less resourced languages*, 1–7. Donostia: SALTMIL.

Borin, Lars 2012. Core vocabulary: A useful but mystical concept in some kinds of linguistics. *Shall we play the festschrift game? Essays on the occasion of Lauri Carlson's 60th birthday*, 53–65. Berlin: Springer.

Borin, Lars 2013. The why and how of measuring linguistic differences. Lars Borin and Anju Saxena (eds), *Approaches to measuring linguistic differences*, 3–26. Berlin: De Gruyter Mouton.

Borin, Lars, Bernard Comrie and Anju Saxena 2013. The Intercontinental Dictionary Series – a rich and principled database for language comparison. Lars Borin and Anju Saxena (eds), *Approaches to measuring linguistic differences*, 285–302. Berlin: De Gruyter Mouton.

Borin, Lars, Devdatt Dubhashi, Markus Forsberg, Richard Johansson, Dimitrios Kokkinakis and Pierre Nugues 2013. Mining semantics for culturomics: towards a knowledge-based approach. *Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing*, 3–10. Association for Computing Machinery.

Borin, Lars and Anju Saxena (eds) 2013. *Approaches to measuring linguistic differences*. Berlin: De Gruyter Mouton.

Borin, Lars, Anju Saxena, Taraka Rama and Bernard Comrie 2014. Linguistic landscaping of South Asia using digital language resources: Genetic vs. areal linguistics. *Ninth international conference on language resources and evaluation (lrec'14)*, 3137–3144.

Bostoen, Koen and Bonny Sands 2012. Clicks in south-western Bantu languages: Contact-induced vs. language-internal lexical change. *Proceedings of the 6th world congress of African linguistics*, Volume 5, 129–140. Köppe Verlag.

Bouchard-Côté, Alexandre, David Hall, Thomas L. Griffiths and Dan Klein 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences* 110 (11): 4224–4229.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard and Quentin D. Atkinson 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337 (6097): 957–960.

Bowern, Claire 2010. Correlates of language change in hunter-gatherer and other 'small' languages. *Language and Linguistics Compass* 4 (8): 665–679.

Brew, Chris and David McKelvie 1996. Word-pair extraction for lexicography. *Proceedings of the second international conference on new methods in language processing*, 45–55. Ankara.

Briscoe, Edward J. (ed.) 2002. *Linguistic evolution through language acquisition*. Cambridge: Cambridge University Press.

Brown, Cecil H., Eric W. Holman and Søren Wichmann 2013. Sound correspondences in the world's languages. *Language* 89 (1): 4–29.

Brown, Cecil H., Eric W. Holman, Søren Wichmann and Viveka Velupillai 2008. Automated classification of the world's languages: A description of the method and preliminary results. *Sprachtypologie und Universalienforschung* 61 (4): 285–308.

Burlak, SA 2014. Languages, DNA, relationship and contacts. *Journal of Language Relationship* 9: 55–67.

Burrow, Thomas H. and Murray B. Emeneau 1984. *A Dravidian etymological dictionary (rev.)*. Oxford: Clarendon Press.

Campbell, Lyle 2003. How to show languages are related: Methods for distant genetic relationship. Brian D. Joseph and Richard D. Janda (eds), *The handbook of historical linguistics*, 262–282. Oxford, UK: Blackwell Publishing.

Campbell, Lyle 2004. *Historical linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Campbell, Lyle 2012. Classification of the indigenous languages of South America. Lyle Campbell and Verónica Grondona (eds), *The indigenous languages of South America*, 59–166. Berlin: De Gruyter Mouton.

Campbell, Lyle and Mauricio J. Mixco 2007. *A glossary of historical linguistics*. University of Utah Press.

Campbell, Lyle and William J. Poser 2008. *Language classification: History and method*. Cambridge University Press.

Cardona, George and Dhanesh Jain 2003. *The Indo-Aryan languages*. London: Routledge.

Cavalli-Sforza, Luigi Luca, Alberto Piazza, Paolo Menozzi and Joanna Mountain 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences* 85 (16): 6002–6006.

Cavnar, William B. and John M. Trenkle 1994. N-gram-based text categorization. *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, 161–175. Las Vegas, USA.

Chang, Will, Chundra Cathcart, David Hall and Andrew Garrett 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91 (1): 194–244.

Chen, Matthew Y. and William S-Y. Wang 1975. Sound change: actuation and implementation. *Language* 51: 255–281.

Christiansen, Chris, Thomas Mailund, Christian NS Pedersen, Martin Randers and Martin Stig Stissing 2006. Fast calculation of the quartet distance between trees of arbitrary degrees. *Algorithms for Molecular Biology* 1, no. 1.

Church, Kenneth Ward and Patrick Hanks 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (1): 22–29.

Cohen, William, Pradeep Ravikumar and Stephen Fienberg 2003. A comparison of string distance metrics for matching names and records. *Kdd workshop on data cleaning, record linkage, and object consolidation*.

Collinge, Neville Edgar 1985. *The laws of Indo-European*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Cooper, Martin C. 2008. Measuring the semantic distance between languages from a statistical analysis of bilingual dictionaries. *Journal of Quantitative Linguistics* 15 (1): 1–33.

Covington, Michael A. 1996. An algorithm to align words for historical comparison. *Computational Linguistics* 22 (4): 481–496.

Croft, William 2000. *Explaining language change: An evolutionary approach*. Pearson Education.

Croft, William 2008. Evolutionary linguistics. *Annual Review of Anthropology* 37 (1): 219–234.

Crowley, Terry and Claire Bowern 2009. *An introduction to historical linguistics*. USA: Oxford University Press.

Cysouw, Michael and Hagen Jung 2007. Cognate identification and alignment using practical orthographies. *Proceedings of ninth meeting of the ACL special interest group in computational morphology and phonology*, 109–116. Association for Computational Linguistics.

Damerau, Fred J. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM* 7 (3): 171–176.

Darwin, Charles 1871. *The descent of man*. London: Murray.

Daume III, Hal 2009. Non-parametric Bayesian areal linguistics. *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, 593–601. Association for Computational Linguistics.

Dawkins, Richard 2006. *The selfish gene*. New York: Oxford university press.

Dediu, Dan and Stephen C Levinson 2013. On the antiquity of language: The reinterpretation of Neandertal linguistic capacities and its consequences. *Frontiers in psychology*, vol. 4.

Dediu, Dan and Stephen C Levinson 2014. The time frame of the emergence of modern language and its implications. Daniel Dor, Chris Knight and Jerome Lewis (eds), *The social origins of language*, 184–195. Oxford: Oxford University Press.

Denham, Tim and Mark Donohue 2012. Reconnecting genes, languages and material culture in Island Southeast Asia: Aphorisms on geography and history. *Language Dynamics and Change* 2: 184–211.

De Oliveira, Paulo Murilo Castro, Adriano O Sousa and Søren Wichmann 2013. On the disintegration of (proto-) languages. *International Journal of the Sociology of Language* 2013 (221): 11–19.

De Oliveira, Paulo Murilo Castro, Dietrich Stauffer, Søren Wichmann and Suzana Moss De Oliveira 2008. A computer simulation of language families. *Journal of Linguistics* 44 (3): 659–675.

Dobson, Annette J., Joseph B. Kruskal, David Sankoff and Leonard J. Savage 1972. The mathematics of glottochronology revisited. *Anthropological Linguistics* 14 (6): 205–212.

Dolgopolsky, Aron B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. Vitalij V. Shevoroshkin and Thomas L. Markey (eds), *Typology, relationship, and time: A collection of papers on language change and relationship by Soviet linguists*, 27–50. Ann Arbor, MI: Karoma.

Donohue, Mark 2012. Typology and Areality. *Language Dynamics and Change* 2 (1): 98–116.

Donohue, Mark and Tim Denham 2011. Languages and genes attest different histories in Island Southeast Asia. *Oceanic Linguistics* 50 (2): 536–542.

Donohue, Mark, Rebecca Hetherington, James McElvenny and Virginia Dawson 2013. World phonotactics database. Department of Linguistics, The Australian National University. http://phonotactics.anu.edu.au.

Downey, Sean S, Brian Hallmark, Murray P Cox, Peter Norquest and J Stephen Lansing 2008. Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction. *Journal of Quantitative Linguistics* 15 (4): 340–369.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in language* 13 (2): 257–292.

Dryer, Matthew S. 2000. Counting genera vs. counting languages. *Linguistic Typology* 4: 334–350.

Dryer, Matthew S. 2011. Genealogical Language List.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson and Russell D. Gray 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473 (7345): 79–82.

Dunn, Michael, Stephen C. Levinson and Eva Lindström 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in island melanesia. *Language* 84 (4): 710–59.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley and Stephen C. Levinson 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309 (5743): 2072–2075.

Dunning, Ted 1994. Statistical identification of language. Technical report CRL MCCS-94-273, Computing Research Lab, New Mexico State University.

Dunning, Ted Emerson 1998. Finding structure in text, genome and other symbolic sequences. Ph.D. diss., University of Sheffield, United Kingdom.

Durbin, Richard, Sean R. Eddy, Anders Krogh and Graeme Mitchison 2002. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press.

Durham, Stanton P. and David Ellis Rogers 1969. An application of computer programming to the reconstruction of a proto-language. *Proceedings of the 1969 conference on computational linguistics*, 1–21. Association for Computational Linguistics.

Durie, Mark and Malcolm Ross (eds) 1996. *The comparative method reviewed: regularity and irregularity in language change*. USA: Oxford University Press.

Dyen, Isidore, Joseph B. Kruskal and Paul Black 1992. An Indo-European classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82 (5): 1–132.

Ebert, Karen 2006. South Asia as a linguistic area. Keith Brown (ed.), *Encyclopedia of languages and linguistics*, 2nd edition, 587–595. Oxford: Elsevier.

Eger, Steffen 2013. Sequence alignment with arbitrary steps and further generalizations, with applications to alignments in linguistics. *Information Sciences* 237 (July): 287–304.

Eger, Steffen and Ineta Sejane 2010. Computing semantic similarity from bilingual dictionaries. *Proceedings of the 10th International Conference on the Statistical Analysis of Textual Data (JADT-2010)*, 1217–1225.

Ellegård, Alvar 1959. Statistical measurement of linguistic relationship. *Language* 35 (2): 131–156.

Ellison, T. Mark and Simon Kirby 2006. Measuring language divergence by intra-lexical comparison. *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, 273–280. Sydney, Australia: Association for Computational Linguistics.

Embleton, Sheila M. 1986. *Statistics in historical linguistics*. Volume 30. Brockmeyer.

Emeneau, Murray B. 1956. India as a linguistic area. *Language* 32: 3–16.

Escalante, Hugo Jair, Thamar Solorio and Manuel Montes-y Gómez 2011. Local histograms of character n-grams for authorship attribution. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, 288–298. Association for Computational Linguistics.

Evans, Steven N., Don Ringe and Tandy Warnow 2006. Inference of divergence times as a statistical inverse problem. Peter Forster and Colin Renfrew (eds), *Phylogenetic methods and the prehistory of languages*, 119–130. Cambridge, UK: McDonald Institute for Archaeological Research, University of Cambridge.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang and Chih-Jen Lin 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research* 9: 1871–1874.

Fellbaum, Christiane 1998. *WordNet: An electronic database*. Cambridge, Massachusetts: MIT Press.

Felsenstein, Joseph 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17 (6): 368–376.

Felsenstein, Joseph 2002. PHYLIP (phylogeny inference package) version 3.6 a3. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Felsenstein, Joseph 2004. *Inferring phylogenies*. Sunderland, Massachusetts: Sinauer Associates.

Fortson, Benjamin W. 2003. An approach to semantic change. Brian D. Joseph and Richard D. Janda (eds), *The handbook of historical linguistics*, 648–666. Wiley Online Library.

Fox, Anthony 1995. *Linguistic reconstruction: An introduction to theory and method*. Oxford: Oxford University Press.

Garrett, Andrew 1999. A new model of Indo-European subgrouping and dispersal. Steve S. Chang, Lily Liaw and Josef Ruppenhofer (eds), *Proceedings of the Twenty-Fifth Annual Meeting of the Berkeley Linguistics Society*, 146–156. Berkeley: Berkeley Linguistic Society.

Gavin, Michael C., Carlos A. Botero, Claire Bowern, Robert K. Colwell, Michael Dunn, Robert R. Dunn, Russell D. Gray, Kathryn R. Kirby, Joe McCarter, Adam Powell, Thiago F. Rangel, John R. Stepp, Michelle Trautwein, Jennifer L. Verdolin and Gregor Yanega 2013. Toward a mechanistic understanding of linguistic diversity. *BioScience* 63 (7): 524–535.

Georgi, Ryan, Fei Xia and William Lewis 2010. Comparing language similarity across genetic and typologically-based groupings. *Proceedings of the 23rd International Conference on Computational Linguistics*, 385–393. Association for Computational Linguistics.

Goodman, Leo A. and William H. Kruskal 1954. Measures of association for cross classifications. *Journal of the American Statistical Association*, pp. 732–764.

Graff, P., Z. Balewski, K. L. Evans, A. Mentzelopoulos, K. Snyder, E. Taliep, M. Tarczon, and X. Wang 2011. The World Lexicon (WOLEX) Corpus. http://www.wolex.org/.

Gravano, Luis, Panagiotis G. Ipeirotis, Hosagrahar Visvesvaraya Jagadish, Nick Koudas, Shanmugauelayut Muthukrishnan, Lauri Pietarinen and Divesh Srivastava 2001. Using q-grams in a DBMS for approximate string processing. *IEEE Data Engineering Bulletin* 24 (4): 28–34.

Gray, Russell D. and Quentin D. Atkinson 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426 (6965): 435–439.

Gray, Russell D., David Bryant and Simon J. Greenhill 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1559): 3923–3933.

Gray, Russell D. and Fiona M. Jordan 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405 (6790): 1052–1055.

Greenberg, Joseph H. 1993. Observations concerning Ringe's "Calculating the factor of chance in language comparison". *Proceedings of the American Philosophical Society* 137 (1): 79–90.

Greenhill, Simon J., Robert Blust and Russell D. Gray 2008. The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics Online* 4: 271–283.

Greenhill, Simon J., Alexei J. Drummond and Russell D. Gray 2010. How accurate and robust are the phylogenetic estimates of Austronesian language relationships? *PloS one* 5 (3): e9573.

Greenhill, Simon J. and Russell D. Gray 2009. Austronesian language phylogenies: Myths and misconceptions about Bayesian computational methods. *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*, pp. 375–397.

Grierson, George A 1903–1927. *A Linguistic Survey of India*. Volume I-XI. Delhi: Motilal Banarasidas.

Grimes, Joseph E. and Frederick B. Agard 1959. Linguistic divergence in Romance. *Language* 35 (4): 598–604.

Gulordava, Kristina and Marco Baroni 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, 67–71. Association for Computational Linguistics.

Gusfield, Dan 1997. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge University Press.

Haas, Mary R. 1958. Algonkian-Ritwan: The end of a controversy. *International Journal of American Linguistics* 24 (3): 159–173.

Hammarström, Harald 2009. Sampling and genealogical coverage in the WALS. *Linguistic Typology* 13 (1): 105–119. Plus 198pp appendix.

Hammarström, Harald 2010. A full-scale test of the language farming dispersal hypothesis. *Diachronica* 27 (2): 197–213.

Hammarström, Harald and Lars Borin 2011. Unsupervised learning of morphology. *Computational Linguistics* 37 (2): 309–350.

Harrison, Sheldon P. 2003. On the limits of the comparative method. Brian D. Joseph and Richard D. Janda (eds), *The handbook of historical linguistics*, 213–243. Wiley Online Library.

Haspelmath, Martin, Matthew S. Dryer, David Gil and Bernard Comrie 2011. *WALS online*. Munich: Max Planck Digital Library. http://wals.info.

Haspelmath, Martin and Uri Tadmor 2009a. The loanword typology project and the world loanword database. Martin Haspelmath and Uri Tadmor (eds), *Loanwords in the world's languages: A comparative handbook*, 1–33. De Gruyter Mouton.

Haspelmath, Martin and Uri Tadmor (eds) 2009b. *Loanwords in the world's languages: A comparative handbook*. De Gruyter Mouton.

Haudricourt, André G. 1961. Richesse en phonèmes et richesse en locateurs. *L'Homme* 1: 5–10.

Hauer, Bradley and Grzegorz Kondrak 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. *Proceedings of 5th international joint conference on natural language processing*, 865–873. Chiang Mai, Thailand: Asian Federation of Natural Language Processing.

Hay, Jennifer and Laurie Bauer 2007. Phoneme inventory size and population size. *Language* 83 (2): 388–400.

Heeringa, Wilbert Jan 2004. Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. diss., University Library Groningen.

Heggarty, Paul 2014. Prehistory through language and archaeology. Claire Bowern and Bethwyn Evans (eds), *The routledge handbook of historical linguistics*, 598–626. Routledge.

Hewson, John 1973. Reconstructing prehistoric languages on the computer: The triumph of the electronic neogrammarian. *Proceedings of the 5th conference on computational linguistics*, Volume 1, 263–273. Association for Computational Linguistics.

Hewson, John 1993. *A computer-generated dictionary of proto-Algonquian*. Hull, Quebec: Canadian Museum of Civilization.

Hewson, John 2010. Sound Change and the Comparative Method: The Science of Historical Reconstruction. Silvia Luraghi and Vít Bubeník (eds), *Continuum companion to historical linguistics*, 39–52. London: Continuum International Publishing Group.

Hock, Hans Heinrich 2010. Typology and universals. Silvia Luraghi and Vít Bubeník (eds), *Continuum companion to historical linguistics*, 59–69. London: Continuum International Publishing Group.

Hock, Hans Henrich 1991. *Principles of historical linguistics*. Berlin: Walter de Gruyter.

Hock, Hans Henrich and Brian D. Joseph 2009. *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*. Volume 218. Berlin: Walter de Gruyter.

Hoenigswald, Henry M. 1963. On the history of the comparative method. *Anthropological Linguistics* 5 (1): 1–11.

Hoenigswald, Henry M. 1973. The comparative method. Thomas Sebeok (ed.), *Current trends in linguistics*, Volume 2, 51–62. Berlin: De Gruyter, Mouton.

Hoenigswald, Henry M. 1987. Language family trees, topological and metrical. Henry M. Hoenigswald and Linda F. Wiener (eds), *Biological metaphor and cladistic classification*, 257–267. London: Pinter, Frances.

Hoenigswald, Henry M. 1990. Descent, perfection and the comparative method since Leibniz. Tullio De Mauro and Lia Formigari (eds), *Leibniz, Humboldt, and the origins of comparativism.*, 119–132. Amsterdam: John Benjamins Publishing Company.

Hoenigswald, Henry M. 1991. Is the "comparative" method general or family-specific? Philip Baldi (ed.), *Pattern of change, change of pattern: Linguistic change and reconstruction methodology*, 183–191. Berlin: Mouton de Gruyter.

Holden, Claire Janaki 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269 (1493): 793–799.

Holden, Claire Janaki and Russell D. Gray 2006. Rapid radiation, borrowing and dialect continua in the Bantu languages. Peter Forster and Colin Renfrew (eds), *Phylogenetic methods and the prehistory of languages*, 19–32. Cambridge, UK: McDonald Institute for Archaeological Research, University of Cambridge.

Holland, Barbara R., Katharina T. Huber, Andreas Dress and Vincent Moulton 2002. δ plots: A tool for analyzing phylogenetic distance data. *Molecular Biology and Evolution* 19 (12): 2051–2059.

Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List and Dmitry Egorov 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology* 52 (6): 841–875.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller and Dik Bakker 2008a. Explorations in automated language classification. *Folia Linguistica* 42 (3-4): 331–354.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller and Dik Bakker 2008b. Advances in automated language classification. Antti Arppe, Kaius Sinnemäki and Urpu Nikanne (eds), *Quantitative investigations in theoretical linguistics*, 40–43. Helsinki: University of Helsinki.

Holman, Eric W, Christian Schulze, Dietrich Stauffer and Søren Wichmann 2007. On the relation between structural diversity and geographical distance among languages: Observations and computer simulations. *Linguistic Typology* 11: 393–421.

Huff, Paul and Deryle Lonsdale 2011. Positing language relationships using ALINE. *Language Dynamics and Change* 1 (1): 128–162.

Huffman, Stephen M. 1998. The genetic classification of languages by n-gram analysis: A computational technique. Ph.D. diss., Georgetown University, Washington, DC, USA. AAI9839491.

Hull, David L. 2001. *Science and selection: Essays on biological evolution and the philosophy of science*. Cambridge, UK: Cambridge University Press.

Hunley, Keith, Michael Dunn, Eva Lindström, Ger Reesink, Angela Terrill, Meghan E Healy, George Koki, Françoise R Friedlaender and Jonathan S Friedlaender 2008. Genetic and linguistic coevolution in Northern Island Melanesia. *PLoS genetics* 4 (10): e1000239.

Huson, Daniel H. and David Bryant 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23 (2): 254–267.

Inkpen, Diana, Oana Frunza and Grzegorz Kondrak 2005. Automatic identification of cognates and false friends in French and English. *Proceedings of the international conference recent advances in natural language processing*, 251–257.

Jäger, Gerhard 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* 3 (2): 245–291.

Jankowska, Magdalena, Vlado Keselj and Evangelos Milios 2012. Relative n-gram signatures: Document visualization at the level of character n-grams. *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, 103–112. IEEE.

Järvelin, Anni, Antti Järvelin and Kalervo Järvelin 2007. s-grams: Defining generalized n-grams for information retrieval. *Information Processing & Management* 43 (4): 1005–1019.

Jordan, Fiona M., Russell D. Gray, Simon J. Greenhill and Ruth Mace 2009. Matrilocal residence is ancestral in Austronesian societies. *Proceedings of the Royal Society B: Biological Sciences* 276 (1664): 1957–1964.

Jurafsky, Daniel and James H. Martin 2000. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Kachru, Braj B., Yamuna Kachru and S. N. Sridhar (eds) 2008. *Language in South Asia*. Cambridge: Cambridge University Press.

Kay, Martin 1964. The logic of cognate recognition in historical linguistics. Technical Report, The Rand Corporation.

Kessler, Brett 1995. Computational dialectology in Irish Gaelic. *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, 60–66. Morgan Kaufmann Publishers Inc.

Kessler, Brett 2001. *The Significance of Word Lists*. Stanford, CA: CSLI Publications.

Kessler, Brett 2007. Word similarity metrics and multilateral comparison. *Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, 6–14. Association for Computational Linguistics.

Kessler, Brett 2008. The mathematical assessment of long-range linguistic relationships. *Language and Linguistics Compass* 2 (5): 821–839.

Klein, Dan, Joseph Smarr, Huy Nguyen and Christopher D Manning 2003. Named entity recognition with character-level models. *Proceedings of the seventh conference on natural language learning at hlt-naacl 2003-volume 4*, 180–183. Association for Computational Linguistics.

Klein, Sheldon, Michael A Kuppin and Kirby A Meives 1969. Monte Carlo simulation of language change in Tikopia & Maori. *Proceedings of the 1969 conference on computational linguistics*, 1–27. Association for Computational Linguistics.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens et al. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. Association for Computational Linguistics.

Kolachina, Sudheer, Taraka Rama and B. Lakshmi Bai 2011. Maximum parsimony method in the subgrouping of Dravidian languages. *QITL* 4: 52–56.

Kondrak, Grzegorz 2000. A new algorithm for the alignment of phonetic sequences. *Proceedings of the first meeting of the North American chapter of the association for computational linguistics*, 288–295.

Kondrak, Grzegorz 2001. Identifying cognates by phonetic and semantic similarity. *Proceedings of the second meeting of the North American chapter of the association for computational linguistics on language technologies*, 1–8. Association for Computational Linguistics.

Kondrak, Grzegorz 2002a. Algorithms for language reconstruction. Ph.D. diss., University of Toronto, Ontario, Canada.

Kondrak, Grzegorz 2002b. Determining recurrent sound correspondences by inducing translation models. *Proceedings of the 19th international conference on computational linguistics-volume 1*, 1–7. Association for Computational Linguistics.

Kondrak, Grzegorz 2004. Combining evidence in cognate identification. *Advances in Artificial Intelligence*, 44–59. Springer.

Kondrak, Grzegorz 2005. N-gram similarity and distance. *String processing and information retrieval*, 115–126. Springer.

Kondrak, Grzegorz 2009a. Identification of cognates and recurrent sound correspondences in word lists. *Traitement Automatique des Langues et Langues Anciennes* 50 (2): 201–235 (October).

Kondrak, Grzegorz 2009b. Identification of cognates and recurrent sound correspondences in word lists. *Traitement Automatique des Langues* 50 (2): 201–235.

Kondrak, Grzegorz and Bonnie Dorr 2006. Automatic identification of confusable drug names. *Artificial Intelligence in Medicine* 36 (1): 29–42.

Kondrak, Grzegorz and Tarek Sherif 2006. Evaluation of several phonetic similarity algorithms on the task of cognate identification. *Proceedings of ACL workshop on linguistic distances*, 43–50. Association for Computational Linguistics.

Kopotev, Mikhail, Lidia Pivovarova, Natalia Kochetkova and Roman Yangarber 2013. Automatic detection of stable grammatical features in n-grams. *NAACL-HLT* 13: 73–81.

Krause, Johannes, Qiaomei Fu, Jeffrey M. Good, Bence Viola, Michael V. Shunkov, Anatoli P. Derevianko and Svante Pääbo 2010. The complete

mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* 464 (7290): 894–897.

Krauss, Michael E. 1992. The world's languages in crisis. *Language* 68 (1): 4–10.

Krishnamurti, Bhadriraju 1978. Areal and lexical diffusion of sound change: Evidence from Dravidian. *Language* 54 (1): 1–20.

Krishnamurti, Bhadriraju 1998. Regularity of sound change through lexical diffusion: A study of s > h > ø in Gondi dialects. *Language Variation and Change* 10: 193–220.

Krishnamurti, Bhadriraju 2003. *The Dravidian languages*. Cambridge Language Surveys. Cambridge: Cambridge University Press.

Krishnamurti, Bhadriraju and Murray Barnson Emeneau 2001. *Comparative Dravidian linguistics: Current perspectives*. Oxford University Press.

Krishnamurti, Bhadriraju, Lincoln Moses and Douglas G. Danforth 1983. Unchanged cognates as a criterion in linguistic subgrouping. *Language* 59 (3): 541–568.

Kroeber, Alfred L and C. D. Chrétien 1937. Quantitative classification of Indo-European languages. *Language* 13 (2): 83–103.

Kroeber, Alfred L and C. D. Chrétien 1939. The statistical technique and Hittite. *Language* 15 (2): 69–71.

Kruskal, Joseph B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* 29 (2): 115–129.

Labov, William 1994. *Principles of linguistic change*. Oxford: Blackwell.

Lakner, Clemens, Paul Van Der Mark, John P Huelsenbeck, Bret Larget and Fredrik Ronquist 2008. Efficiency of markov chain monte carlo tree proposals in bayesian phylogenetics. *Systematic biology* 57 (1): 86–103.

Lees, Robert B. 1953. The basis of glottochronology. *Language* 29 (2): 113–127.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady*, Volume 10, 707.

Lewis, M. Paul, Gary F. Simons and Charles D. Fennig (eds) 2013. *Ethnologue: Languages of the world*. Seventeenth edition. Dallas, TX: SIL International. Online version: `http://www.ethnologue.com`.

Lewis, Paul M. (ed.) 2009. *Ethnologue: Languages of the world*. Sixteenth edition. Dallas, TX, USA: SIL International.

Lewis, William D. and Fei Xia 2010. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing* 25 (3): 303–319.

Liang, Percy, Ben Taskar and Dan Klein 2006. Alignment by agreement. *Proceedings of the main conference on human language technology conference of the north american chapter of the association of computational linguistics*, 104–111. Association for Computational Linguistics.

Lin, Dekang 1998. An information-theoretic definition of similarity. *Proceedings of the 15th international conference on machine learning*, Volume 1, 296–304.

List, Johann-Mattis 2012. LexStat: Automatic detection of cognates in multilingual wordlists. *Proceedings of the EACL 2012 joint workshop of LINGVIS & UNCLH*, 117–125. Avignon, France: Association for Computational Linguistics.

List, Johann-Mattis and Steven Moran 2013. An open source toolkit for quantitative historical linguistics. *Proceedings of the 51st annual meeting of the association for computational linguistics: System demonstrations*, 13–18. Sofia, Bulgaria: Association for Computational Linguistics.

List, Johann-Mattis and Jelena Prokić 2014. A benchmark database of phonetic alignments in historical linguistics and dialectology. N. C. (. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (eds), *Proceedings of the ninth international conference on language resources and evaluation*, 288–294. European Language Resources Association (ELRA).

Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini and Chris Watkins 2002. Text classification using string kernels. *The Journal of Machine Learning Research* 2: 419–444.

Lohr, Marisa 1998. Methods for the genetic classification of languages. Ph.D. diss., University of Cambridge.

Lowe, John B. and Martine Mazaudon 1994. The reconstruction engine: A computer implementation of the comparative method. *Computational Linguistics* 20 (3): 381–417.

Luján, Eugenio R. 2010. Semantic change. Silvia Luraghi and Vít Bubeník (eds), *Continuum companion to historical linguistics*, 286–310. London: Continuum International Publishing Group.

Maddieson, Ian 1984. *Pattern of Sounds*. Cambridge, UK: Cambridge University Press.

Maddieson, Ian and Kristin Precoda 1990a. UPSID-PC. *The UCLA Phonological Segment Inventory Database*.

Maddieson, Ian and Kristin Precoda 1990b. Updating UPSID. *UCLA working papers in phonetics*, Volume 74, 104–111. Department of Linguistics, UCLA.

Mallory, James P 2013. Twenty-first century clouds over indo-european homelands. *Journal of Language Relationship* 9: 145–154.

Mantel, Nathan 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27 (2 Part 1): 209–220.

Masica, Colin P. 1976. *Defining a linguistic area: South Asia*. Chicago: University of Chicago Press.

Masica, Colin P. 1993. *The Indo-Aryan languages*. Cambridge Language Surveys. Cambridge: Cambridge University Press.

Matisoff, James A. 2003. *Handbook of Proto-Tibeto-Burman. system and philosophy of Sino-Tibetan reconstruction*. Berkeley: University of California Press.

McCullough, Peter and John A Nelder 1989. *Generalized linear models*. London: Chapman & Hall.

McMahon, April, Paul Heggarty, Robert McMahon and Warren Maguire 2007. The sound patterns of Englishes: representing phonetic similarity. *English Language and Linguistics* 11 (01): 113–142.

McMahon, April M. S. and Robert McMahon 2005. *Language classification by numbers*. USA: Oxford University Press.

McMahon, April M.S. and Robert McMahon 2007. Language families and quantitative methods in South Asia and elsewhere. Michael D. Petraglia and Bridget Allchin (eds), *The Evolution and History of Human Populations in South Asia*, Vertebrate Paleobiology and Paleoanthropology Series, 363–384. Netherlands: Springer.

Meillet, Antoine 1967. *The comparative method in historical linguistics*. Paris: Librairie Honoré Champion. Translated by Gordon B. Ford.

Melamed, Dan I. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics* 25 (1): 107–130.

Metcalf, George J. 1974. The Indo-European hypothesis in the sixteenth and seventeenth centuries. Dell Hymes (ed.), *Studies in the history of linguistics: Traditions and paradigms*, 233–257. Bloomington: Indiana University Press.

Moran, Steven 2012. Using linked data to create a typological knowledge base. Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds), *Linked data in linguistics: Representing and connecting language data and language metadata*, 129–138. Heidelberg: Springer. doi:10.1007/978-3-642-28249-2_13.

Moschitti, Alessandro, Qi Ju and Richard Johansson 2012. Modeling topic dependencies in hierarchical text categorization. *Proceedings of the 50th*

annual meeting of the association for computational linguistics: Long papers-volume 1, 759–767. Association for Computational Linguistics.

Nakhleh, Luay, Tandy Warnow, Don Ringe and Steven N. Evans 2005. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. *Transactions of the Philological Society* 103 (2): 171–192.

Nakleh, Luay, Don Ringe and Tandy Warnow 2005. Perfect phylogentic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81 (2): 382–420.

Nakov, Preslav and Jörg Tiedemann 2012. Combining word-level and character-level models for machine translation between closely-related languages. *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, 301–305. Association for Computational Linguistics.

Needleman, Saul B. and Christian D. Wunsch 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48 (3): 443–453.

Nerbonne, John, Wilbert Heeringa and Peter Kleiweg 1999. Edit distance and dialect proximity. David Sankoff and Joseph Kruskal (eds), *Time warps, string edits and macromolecules: The theory and practice of sequence comparison*, 2, V–XV. Stanford, CA: CSLI publications.

Nerbonne, John and Erhard Hinrichs 2006. Linguistic distances. *Proceedings of the workshop on linguistic distances*, 1–6. Association for Computational Linguistics.

Nettle, Daniel 1995. Segmental inventory size, word length, and communicative efficiency. *Linguistics* 33 (2): 359–367.

Nettle, Daniel 1998. Explaining global patterns of language diversity. *Journal of Anthropological Archaeology* 17 (4): 354–374.

Nettle, Daniel 1999a. *Linguistic diversity*. Oxford: Oxford University Press.

Nettle, Daniel 1999b. Using Social Impact Theory to simulate language change. *Lingua* 108 (2-3): 95–117.

Nichols, Johanna 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Nichols, Johanna 1995. Diachronically stable structural features. Henning Andersen (ed.), *Historical linguistics 1993. Selected papers from the 11th international conference on historical linguistics*, 337–355. Amsterdam/Philadelphia: John Benjamins.

Nichols, Johanna 1996. The comparative method as heuristic. Mark Durie and Malcom Ross (eds), *The comparative method revisited: Regularity and*

irregularity in language change, 39–71. New York: Oxford University Press.

Nichols, Johanna and Tandy Warnow 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2 (5): 760–820.

Niyogi, Partha 2006. *The computational nature of language learning and evolution*. Volume 43 of *Current studies in linguistics*. Cambridge: MIT Press.

Nordhoff, Sebastian and Harald Hammarström 2011. Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. *Proceedings of the first international workshop on linked science*, Volume 783.

Nordhoff, Sebastian and Harald Hammarström 2012. Glottolog/Langdoc: Increasing the visibility of grey literature for low-density languages. *Language resources and evaluation conference*, 3289–3294.

Nowak, Martin A., Natalia L. Komarova and Partha Niyogi 2002. Computational and evolutionary aspects of language. *Nature* 417 (6889): 611–617.

Och, Franz Josef and Hermann Ney 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29 (1): 19–51.

Oswalt, Robert L. 1971. Towards the construction of a standard lexicostatistic list. *Anthropological Linguistics* 13 (9): 421–434.

Pagel, Mark 1999. Inferring the historical patterns of biological evolution. *Nature* 401 (6756): 877–884.

Pakendorf, Brigitte 2014. Coevolution of languages and genes. *Current opinion in genetics & development* 29: 39–44.

Petroni, Filippo and Maurizio Serva 2010. Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications* 389 (11): 2280–2283.

Petroni, Filippo and Maurizio Serva 2011. Automated word stability and language phylogeny. *Journal of Quantitative Linguistics* 18 (1): 53–62.

Pettersson, Eva, Beáta B. Megyesi and Jörg Tiedemann 2013. An SMT approach to automatic annotation of historical text. *Proceedings of the Nodalida Workshop on Computational Historical Linguistics*. Oslo, Norway: NEALT.

Piotrowski, Michael 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies* 5 (2): 1–157.

Plank, Frans 1998. The co-variation of phonology with morphology and syntax: A hopeful history. *Linguistic Typology* 2 (2): 195–230.

Polyakov, Vladimir N., Valery D. Solovyev, Søren Wichmann and Oleg Belyaev 2009. Using WALS and Jazyki Mira. *Linguistic Typology* 13 (1): 137–167.

Pompei, Simone, Vittorio Loreto and Francesca Tria 2011. On the accuracy of language trees. *PloS one* 6 (6): e20109.

Porta, Jordi and José-Luis Sancho 2014. Using maximum entropy models to discriminate between similar languages and varieties. *COLING 2014*, p. 120.

Poser, William J. and Lyle Campbell 1992. Indo-European practice and historical methodology. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, Volume 18, 214–236.

Quintana-Murci, Lluís, Hélène Quach, Christine Harmant, Francesca Luca, Blandine Massonnet, Etienne Patin, Lucas Sica, Patrick Mouguiama-Daouda, David Comas, Shay Tzur, Oleg Balanovsky, Kenneth K. Kidd, Judith R. Kidd, Lolke van der Veen, Jean-Marie Hombert, Antoine Gessain, Paul Verdu, Alain Froment, Serge Bahuchet, Evelyne Heyer, Jean Dausset, Antonio Salas and Doron M. Behar 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter–gatherers and Bantu-speaking farmers. *Proceedings of the National Academy of Sciences* 105 (5): 1596–1601.

R Core Team 2012. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rama, Taraka 2013. Phonotactic diversity predicts the time depth of the world's language families. *PloS one* 8 (5): e63238.

Rama, Taraka 2015. Automatic cognate identification with gap-weighted string subsequences. *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies.*, 1227–1231.

Rama, Taraka and Lars Borin 2013. N-gram approaches to the historical dynamics of basic vocabulary. *Journal of Quantitative Linguistics* 21 (1): 50–64.

Rama, Taraka and Lars Borin 2015. Comparative evaluation of string similarity measures for automatic language classification. Ján Mačutek and George K. Mikros (eds), *Sequences in language and text*, 203–231. Walter de Gruyter.

Rama, Taraka, Prasant Kolachina and Sudheer Kolachina 2013. Two methods for automatic identification of cognates. *QITL* 5: 76.

Rama, Taraka and Prasanth Kolachina 2012. How good are typological distances for determining genealogical relationships among languages? *COLING (posters)*, 975–984.

Rama, Taraka and Sudheer Kolachina 2013. Distance-based phylogenetic inference algorithms in the subgrouping of Dravidian languages. Lars Borin and Anju Saxena (eds), *Approaches to measuring linguistic differences*, 141–174. Berlin: De Gruyter Mouton.

Rama, Taraka, Sudheer Kolachina and B. Lakshmi Bai 2009. Quantitative methods for phylogenetic inference in historical linguistics: An experimental case study of South Central Dravidian. *Indian Linguistics* 70: 265–282.

Rama, Taraka and Anil Kumar Singh 2009. From bag of languages to family trees from noisy corpus. *Proceedings of the International Conference RANLP-2009*, 355–359. Borovets, Bulgaria: Association for Computational Linguistics.

Raman, Anand and Jon Patrick 1997. Linguistic similarity measures using the minimum message length principle. Roger Blench and Matthew Spriggs (eds), *Archaeology and language I: Theoretical and methodological orientations*, 262–279. Routledge.

Rankin, Robert L. 2003. The comparative method. Brian D. Joseph and Richard D. Janda (eds), *The handbook of historical linguistics*, 199–212. Wiley Online Library.

Ravi, Sujith and Kevin Knight 2008. Attacking decipherment problems optimally with low-order n-gram models. *Proceedings of the conference on empirical methods in natural language processing*, 812–819. Association for Computational Linguistics.

Reddy, Sravana and Kevin Knight 2011. What we know about the Voynich manuscript. *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, 78–86. Association for Computational Linguistics.

Reich, David, Nick Patterson, Desmond Campbell, Arti Tandon, Stéphane Mazieres, Nicolas Ray, Maria V Parra, Winston Rojas, Constanza Duque, Natalia Mesa et al. 2012. Reconstructing native American population history. *Nature* 488 (7411): 370–374.

Renfrew, Colin, April M. S. McMahon and Robert Lawrence Trask 2000. *Time depth in historical linguistics*. McDonald Institute for Archaeological Research.

Ringe, Don, Tandy Warnow and Ann Taylor 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100

(1): 59–129.

Ringe, Donald 2006. *From Proto-Indo-European to Proto-Germanic: A linguistic history of English*. Volume 1. Oxford University Press.

Ringe, Donald A. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82 (1): 1–110.

Ritt, Nikolaus 2004. *Selfish sounds and linguistic evolution: A Darwinian approach to language change*. Cambridge, UK: Cambridge University Press.

Robinson, D. and L. Foulds 1979. Comparison of weighted labelled trees. *Combinatorial mathematics VI*, pp. 119–126.

Ronquist, Fredrik and John P Huelsenbeck 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19 (12): 1572–1574.

Ross, Alan S. C. 1950. Philological probability problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 19–59.

Ruhlen, Merritt 1991. *A guide to the world's languages: Classification*. Volume 1. Stanford, California: Stanford University Press.

Saitou, Naruya and Masatoshi Nei 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4 (4): 406–425.

Sankoff, David 1969. Historical linguistics as stochastic process. Ph.D. diss., McGill University.

Sapir, Edward 1916. *Time perspective in aboriginal American culture: A study in method*. Anthropological Linguistics no. 13. Ottawa: Government Printing Bureau. Geological Survey of Canada Memoir 90.

Saussure, Ferdinand De 1879. *Mémoire sur le système primitif des voyelles dans les langues indo-européennes*. Leipsick: BG Teubner.

Saxena, Anju 2011. Towards empirical classification of Kinnauri varieties. Peter K. Austin, Oliver Bond, David Nathan and Lutz Marten (eds), *Proceedings of conference on language documentation & linguistic theory 3*. London: SOAS.

Saxena, Anju and Lars Borin 2011. Dialect classification in the Himalayas: A computational approach. *NODALIDA 2011 conference proceedings*, 307–310. Riga: NEALT.

Saxena, Anju and Lars Borin 2013. Carving Tibeto-Kanauri by its joints: Using basic vocabulary lists for genetic grouping of languages. Lars Borin and Anju Saxena (eds), *Approaches to measuring linguistic differences*,

Trends in Linguistics. Studies and Monographs no. 265, 175–198. Berlin: De Gruyter Mouton.

Schenker, Alexander M 1995. *The dawn of Slavic: An introduction to Slavic philology*. New Haven: Yale University Press.

Schmidt, Johannes 1872. *Die Verwantschaftsverhältnisse der indogermanischen Sprachen*. Weimar: Böhlau.

Sellers, Peter H 1974. On the theory and computation of evolutionary distances. *SIAM Journal on Applied Mathematics* 26 (4): 787–793.

Serva, Maurizio and Filippo Petroni 2008. Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)* 81: 68005.

Shawe-Taylor, John and Nello Cristianini 2004. *Kernel methods for pattern analysis*. Cambridge university press.

Sheskin, David J 2003. *Handbook of parametric and nonparametric statistical procedures*. Chapman & Hall/CRC Press.

Sicoli, Mark A. and Gary Holton 2014. Linguistic phylogenies support back-migration from Beringia to Asia. *PLoS ONE* 9 (3): e91722 (03).

Sidwell, Paul 2010. The austroasiatic central riverine hypothesis. *Journal of Language Relationship (Moscow)* 4: 117–34.

Sidwell, Paul and Roger Blench 2011. The Austroasiatic Urheimat: the Southeastern Riverine Hypothesis. Nicholas J Enfield (ed.), *The dynamics of human diversity*, 1–30. Canberra: Pacific Linguistics.

Simard, Michel, George F. Foster and Pierre Isabelle 1993. Using cognates to align sentences in bilingual corpora. *Proceedings of the 1993 conference of the centre for advanced studies on collaborative research: distributed computing*, Volume 2, 1071–1082. IBM Press.

Singh, Anil Kumar 2006. Study of some distance measures for language and encoding identification. *Proceeding of ACL 2006 workshop on linguistic distances*. Sydney, Australia: Association for Computational Linguistics.

Singh, Anil Kumar and Harshit Surana 2007. Can corpus based measures be used for comparative study of languages? *Proceedings of ninth meeting of the ACL special interest group in computational morphology and phonology*, 40–47. Association for Computational Linguistics.

Smith, Raoul N. 1969. Automatic simulation of historical change. *Proceedings of the 1969 conference on computational linguistics*, 1–14. Association for Computational Linguistics.

Smyth, Bill 2003. *Computing patterns in strings*. Pearson Education.

Sokal, Robert R 1988. Genetic, geographic, and linguistic distances in Europe. *Proceedings of the National Academy of Sciences* 85 (5): 1722–1726.

Sokal, Robert R and Charles D Michener 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38: 1409–1438.

Southworth, Franklin C. 1964. Family-tree diagrams. *Language* 40 (4): 557–565.

Starostin, Sergei A. 1991. *Altajskaja Problema i Proisxozhdenie Japonskogo Jazyka [The Altaic Problem and the Origin of the Japanese Language]*. Moscow: Nauka Publishers.

Steel, Mike A. and David Penny 1993. Distributions of tree comparison metrics—some new results. *Systematic Biology* 42 (2): 126–141.

Steiner, Lydia, Peter F. Stadler and Michael Cysouw 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change* 1 (1): 89–127.

Stoneking, Mark 2006. Disentangling genes, geography, and language. *Languages and genes*.

Subbarao, Karumuri Y. 2008. Typological characteristics of South Asian languages. Braj B. Kachru, Yamuna Kachru and S. N. Sridhar (eds), *Language in South Asia*, 49–78. Cambridge: Cambridge University Press.

Swadesh, Morris 1948. The time value of linguistic diversity. Paper presented at the Viking Fund Supper Conference for Anthropologists, 1948.

Swadesh, Morris 1950. Salish internal relationships. *International Journal of American Linguistics* 16 (4): 157–167.

Swadesh, Morris 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society* 96 (4): 452–463.

Swadesh, Morris 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21 (2): 121–137.

Swadesh, Morris 1959. The mesh principle in comparative linguistics. *Anthropological linguistics* 1 (2): 7–14.

Swadesh, Morris 1971. *The origin and diversification of language*. Joel Sherzer (ed.). London: Routledge & Paul, Kegan.

Tadmor, Uri, Martin Haspelmath and Bradley Taylor 2010. Borrowability and the notion of basic vocabulary. *Diachronica* 27 (2): 226–246.

Tahmasebi, Nina 2013. Models and algorithms for automatic detection of language evolution. Towards finding and interpreting of content in long-term archives. Ph.D. diss., Leibniz Universität, Hannover.

Tahmasebi, Nina and Thomas Risse 2013. The role of language evolution in digital archives. *Proc. of 3rd international workshop on semantic digital archives*.

Tate, Robert F. 1954. Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of mathematical statistics* 25 (3): 603–607.

Thurgood, Graham and Randy LaPolla (eds) 2003. *The Sino-Tibetan languages*. London: Routledge.

Tiedemann, Jörg 1999. Automatic construction of weighted string similarity measures. *Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, 213–219.

Trask, Robert Lawrence 1996. *Historical linguistics*. London: Oxford University Press.

Trask, Robert Lawrence 2000. *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh University Press.

Trudgill, Peter 2002. Linguistic and social typology. *The handbook of language variation and change*, 702–728. Oxford: Blackwell.

Turner, Sir Ralph L. 1964. *A comparative dictionary of the Indo-Aryan languages*. Oxford: Oxford University Press.

Vigilant, Linda, Mark Stoneking, Henry Harpending, Kristen Hawkes and Alan C. Wilson 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253 (5027): 1503–1507.

Walker, Robert S., Søren Wichmann, Thomas Mailund and Curtis J. Atkisson 2012. Cultural phylogenetics of the Tupi language family in Lowland South America. *PloS one* 7 (4): e35025.

Wang, William S-Y. 1969. Project DOC: Its methodological basis. *Proceedings of the 1969 conference on computational linguistics*, 1–22. Association for Computational Linguistics.

Warnow, Tandy 1997. Mathematical approaches to comparative linguistics. *Proceedings of the National Academy of Sciences* 94 (13): 6585–6590.

Wettig, Hannes 2013. Probabilistic, information-theoretic models for etymological alignment. Ph.D. diss., University of Helsinki, Finland.

Wichmann, Søren, Eric W Holman and Cecil H Brown 2010. Sound symbolism in basic vocabulary. *Entropy* 12 (4): 844–858.

Wichmann, Søren 2008. The emerging field of language dynamics. *Language and Linguistics Compass* 2 (3): 442–455.

Wichmann, Søren 2010a. Internal language classification. Silvia Luraghi and Vít Bubeník (eds), *Continuum companion to historical linguistics*, 70–88. London: Continuum International Publishing Group.

Wichmann, Søren 2010b. Neolithic linguistics. Manuscript in possesssion of the author.

Wichmann, Søren 2013. A classification of Papuan languages. *Language and Linguistics in Melanesia*, pp. 313–386.

Wichmann, Søren and Jeff Good 2011. Editorial statement. *Language Dynamics and Change* 1 (1): 1–2.

Wichmann, Søren and Eric W. Holman 2009a. Population size and rates of language change. *Human Biology* 81 (2-3): 259–274.

Wichmann, Søren and Eric W. Holman 2009b. *Assessing temporal stability for linguistic typological features*. München: LINCOM Europa.

Wichmann, Søren, Eric W. Holman, Dik Bakker and Cecil H. Brown 2010a. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications* 389: 3632–3639.

Wichmann, Søren, Eric W. Holman and Johann-Mattis List 2013. The automated classification of the world's languages: Can it go deeper? Presented at QITL-5, Leuven, Belgium.

Wichmann, Søren, Eric W. Holman, Taraka Rama and Robert S. Walker 2011a. Correlates of reticulation in linguistic phylogenies. *Language Dynamics and Change* 1 (2): 205–240.

Wichmann, Søren, André Müller and Viveka Velupillai 2010. Homelands of the world's language families: A quantitative approach. *Diachronica* 27 (2): 247–276.

Wichmann, Søren, André Müller, Viveka Velupillai, Cecil H. Brown, Eric W. Holman, Pamela Brown, Sebastian Sauppe, Oleg Belyaev, Matthias Urban, Zarina Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer, David Beck and Helen Geyer 2010b. The ASJP Database (version 13).

Wichmann, Søren, André Müller, Viveka Velupillai, Cecil H. Brown, Eric W. Holman, Pamela Brown, Matthias Urban, Sebastian Sauppe, Oleg Belyaev, Zarina Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer, David Beck and Helen Geyer 2010c. The ASJP Database (version 12).

Wichmann, Søren, André Müller, Viveka Velupillai, Annkathrin Wett, Cecil H. Brown, Zarina Molochieva, Sebastian Sauppe, Eric W. Holman, Pamela Brown, Julia Bishoffberger, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Helen

Geyer, David Beck, Evgenia Korovina, Pattie Epps, Pilar Valenzuela, Anthony Grant and Harald Hammarström 2011b. The ASJP Database (version 14). http://email.eva.mpg.de/ wichmann/listss14.zip.

Wichmann, Søren and Taraka Rama 2014. Jackknifing the black sheep: ASJP classification performance and Austronesian. Submitted to the proceedings of the symposium "Let's talk about trees", National Museum of Ethnology, Osaka, Febr. 9-10, 2013.

Wichmann, Søren, Taraka Rama and Eric W. Holman 2011. Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15: 177–198.

Wichmann, Søren and Apriar Saunders 2007. How to use typological databases in historical linguistic research. *Diachronica* 24 (2): 373–404.

Wieling, Martijn, Jelena Prokić and John Nerbonne 2009. Evaluating the pairwise string alignment of pronunciations. *Proceedings of the EACL 2009 workshop on language technology and resources for cultural heritage, social sciences, humanities, and education*, 26–34. Association for Computational Linguistics.

Wiersma, Wybo, John Nerbonne and Timo Lauttamus 2011. Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing* 26 (1): 107–124.

Wilks, Yorick, Brian M. Slator and Louise M. Guthrie 1996. *Electric words: dictionaries, computers, and meanings*. Cambridge, MA: MIT Press.

Yang, Ziheng 2014. *Molecular evolution: A statistical approach*. Oxford: Oxford University Press.

Yarowsky, David 1995. Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of ACL*, 189–196.

Yule, Henry and Arthur Coke Burnell 1996. *Hobson-Jobson: The Anglo-Indian dictionary*. Ware: Wordsworth Reference.

A Appendix to publication I

The online supplementary information to "Phonological diversity, word length, and population sizes across languages: The ASJP evidence" is available at the following URL:

http://dx.doi.org/10.1515/LITY.2011.013_suppl_1

B SUPPLEMENTARY INFORMATION TO PHONOTACTIC DIVERSITY

This chapter contains the supplementary information (tables and figures) referred in publication IV (chapter 10).

B.1 Data

Table B.1 presents the details of the calibration points used in the experiments.

Language group	NOL	CD	Type	Family name	MOS	Geographic area
Benue-Congo	404	6500	A	Niger-Congo	AGR	Africa
Brythonic	2	1450	H	Indo-European	AGR	Eurasia
Central SouthernAfrica Khoisan	7	2000	A	Khoisan	PAS	Africa
Cham	2	529	H	Austronesian	AGR	Oceania
Chamic	7	1550	H	Austronesian	AGR	Oceania
Chinese	7	2000	H	Sino-Tibetan	AGR	Eurasia
Cholan	5	1600	E	Mayan	AGR	Americas
Common Turkic	50	1419	H	Altaic	AGR	Eurasia
Czech-Slovak	2	1050	E	Indo-European	AGR	Eurasia
Dardic	22	3550	A	Indo-European	AGR	Eurasia
East Polynesian	11	950	A	Austronesian	AGR	Oceania
East Slavic	4	760	H	Indo-European	AGR	Eurasia

Eastern Malayo-Polynesian	472	3350	A	Austronesian	AGR	Oceania
English-Frisian	4	1550	H	Indo-European	AGR	Eurasia
Ethiopian Semitic	18	2450	E	Afro-Asiatic	AGR	Africa
Ga-Dangme	2	600	AH	Niger-Congo	AGR	Africa
Germanic	30	2100	H	Indo-European	AGR	Eurasia
Goidelic	3	1050	E	Indo-European	AGR	Eurasia
Hmong-Mien	14	2500	E	Hmong-Mein	AGR	Eurasia
Indo-Aryan	93	3900	A	Indo-European	AGR	Eurasia
Indo-European	218	5500	A	Indo-European	AGR	Eurasia
Indo-Iranian	147	4400	A	Indo-European	AGR	Eurasia
Inuit	4	800	A	Eskimo-Aleut	PAS	Americas
Iranian	54	3900	A	Indo-European	AGR	Eurasia
Italo-Western Romance	12	1524	H	Indo-European	AGR	Eurasia
Ket-Yugh	2	1300	H	Yeniseian	PAS	Eurasia
Maa	3	600	H	Nilo-Saharan	AGR	Africa
Malagasy	20	1350	A	Austronesian	AGR	Oceania
Malayo-Chamic	30	2400	A	Austronesian	AGR	Oceania
Malayo-Polynesian	954	4250	A	Austronesian	AGR	Oceania
Maltese-Maghreb Arabic	3	910	H	Afro-Asiatic	AGR	Africa
Mississippi Valley Siouan	9	2475	A	Siouan	PAS	Americas
Mongolic	8	750	H	Altaic	AGR	Eurasia

Northern Roglai Tsat	2	1000	H	Austronesian	AGR	Oceania
Ongamo-Maa	4	1150	A	Nilo-Saharan	AGR	Africa
Oromo	6	460	E	Afro-Asiatic	AGR	Africa
Pama-Nyungan	122	4500	A	Australian	PAS	Oceania
Romance	14	1729	H	Indo-European	AGR	Eurasia
Romani	26	650	H	Indo-European	AGR	Eurasia
Sami	6	1750	A	Uralic	PAS	Eurasia
Scandinavian	7	1100	E	Indo-European	AGR	Eurasia
Slavic	16	1450	H	Indo-European	AGR	Eurasia
Sorbian	3	450	E	Indo-European	AGR	Eurasia
Southern Nilotic	11	2500	A	Nilo-Saharan	AGR	Africa
Southern Songhai	6	550	H	Nilo-Saharan	AGR	Africa
Southwest Tungusic	3	236	H	Altaic	AGR	Eurasia
Swahili	10	1200	AH	Niger-Congo	AGR	Africa
Temotu	9	3200	A	Austronesian	AGR	Oceania
Tupi-Guarani	10	1750	AH	Tupi	AGR	Americas
Turkic	51	2500	AH	Altaic	AGR	Eurasia
Wakashan	5	2500	A	Wakashan	PAS	Americas
Western Turkic	11	900	H	Altaic	AGR	Eurasia

Table B.1: NOL stands for number of languages, CD for calibration dates and MOS for mode of subsistence. In column "Type": 'A' is archaeological, 'AH' is archaeological and historical, 'H' is historical and 'E' is epigraphic calibration points. In column MOS: 'AGR' is agricultural and 'PAS' is foraging and pastoral.

B.2 Diagnostic plots

In this section, we present the four standard diagnostic plots for a linear regression analysis. Each plot has four sub-plots. The sub-plots from left-to-right in each row are summarized as followed:

- The scatter plot of the jackknife deviance residuals vs the predicted value on a log scale.
- The residuals fitted against a standard normal distribution for testing the normality assumption of the residuals.
- A scatterplot showing the Cook's statistic vs. the leverage of each observation. Cook statistic suggests any points which influence the estimation of the regression parameters through a jackknifing procedure. The leverage points are those observations whose omission influences the error value.
- A case plot of the Cook's statistic.

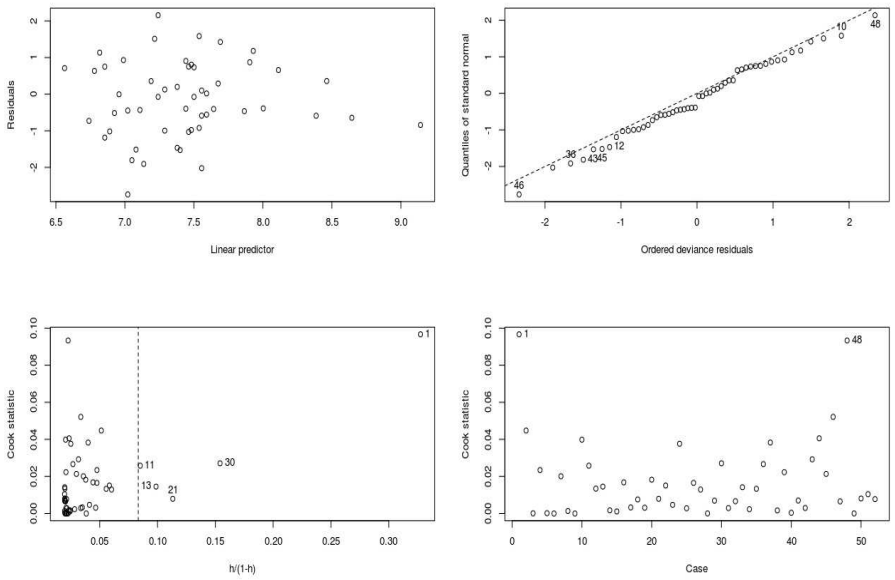


Figure B.1: Diagnostic plots for 1-grams

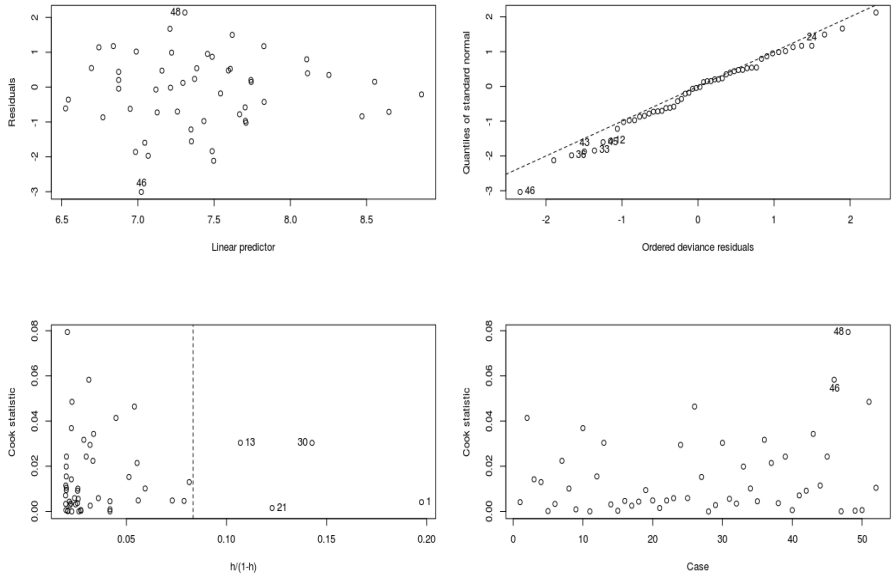


Figure B.2: Diagnostic plots for 2-grams

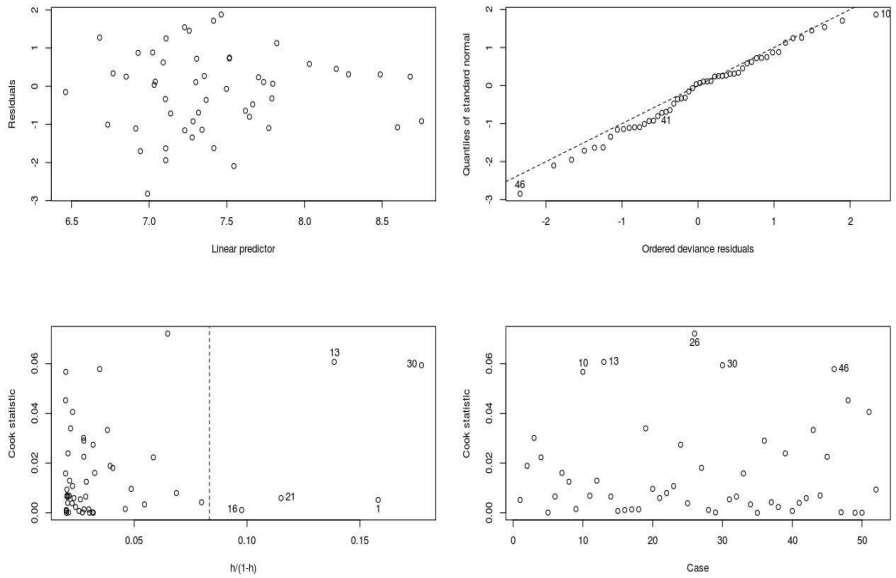


Figure B.3: Diagnostic plots for 3-grams

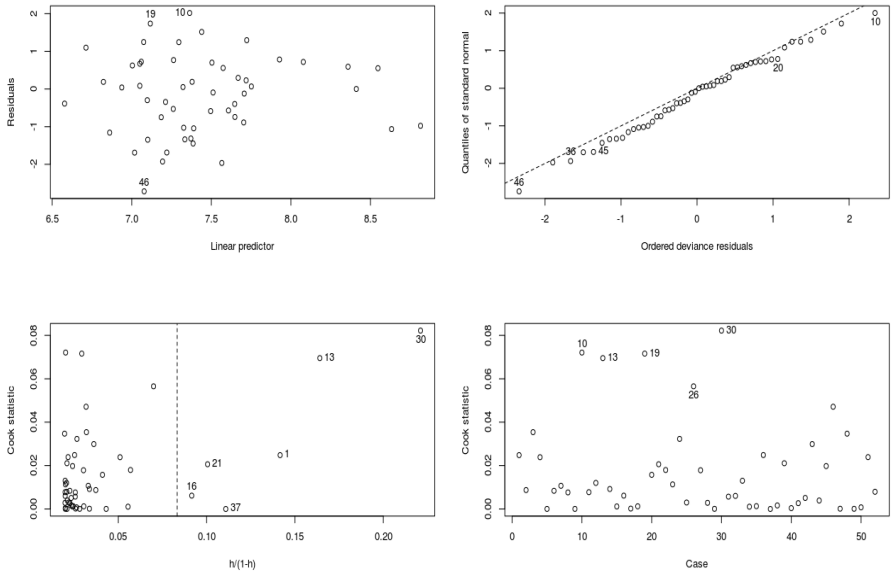


Figure B.4: Diagnostic plots for 4-grams

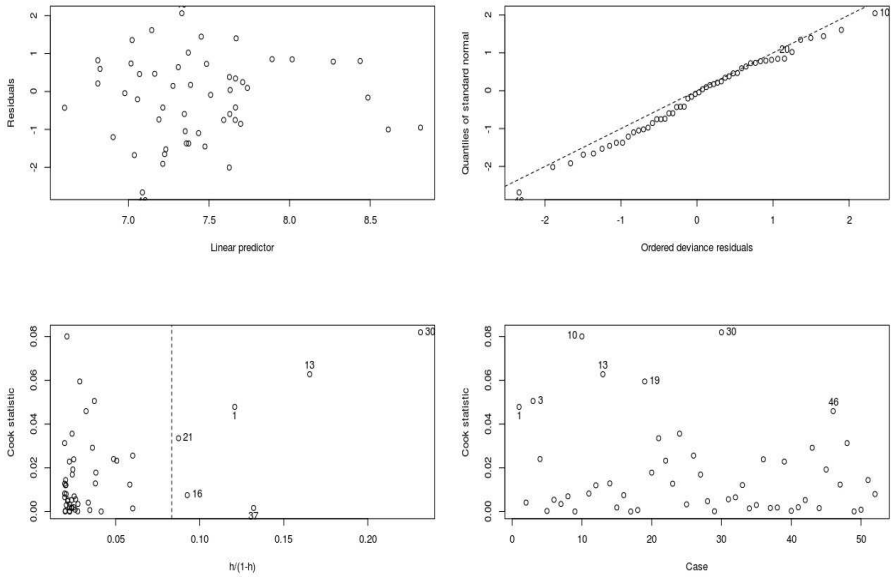


Figure B.5: Diagnostic plots for 5-grams

B.3 Dates of the world's languages

The following tables present the predicted dates for the world languages from Africa, Eurasia, Pacific, North and Middle America, and South America. NOL and CD are Number of Languages and Calibration Date.

Language group	NOL	ASJP date	3-grams date	CD
Afro-Asiatic	255	6016	5769	5915
Berber	23	1733	2220	1933
Eastern	3	1697	1159	1476
Northern	15	1158	1750	1401
Tamasheq	4	556	1208	823
Chadic	98	4826	4214	4575
Biu-Mandara	45	4457	3299	3982
Masa	8	1649	1526	1599
West	40	4099	2943	3625
Cushitic	61	4734	3421	4196
Central	8	1686	1493	1607
East	46	3045	3013	3032
South	6	2308	1522	1986
Omotic	31	4968	2622	4006
North	28	3137	2481	2868
South	3	1963	1108	1612
Semitic	40	3301	3234	3274
Central	18	2638	2405	2542
South	22	3804	2557	3293
Khoisan	17	14592	1863	9373
SouthernAfrica	15	5271	1676	3797
Central	7	3143	1223	2356
Northern	3	1846	873	1447
Southern	5	4344	936	2947
Niger-Congo	679	6227	6889	6498
Atlantic-Congo	594	6525	6672	6585
Atlantic	32	2582	2773	2660
Northern	21	6480	2389	4803
Southern	10	5055	1712	3684
Ijoid	34	4546	1831	3433
Volta-Congo	528	5484	6476	5891
Benue-Congo	404	4940	5887	5328
Dogon	10	2202	1471	1902

Kru	5	2317	1012	1782
Kwa	35	4212	2773	3622
Kordofanian	20	4861	2407	3855
Heiban	11	2521	1789	2221
Katla	2	2269	1086	1784
Talodi	6	4658	1507	3366
Mande	64	3417	2520	3049
Eastern	17	1905	1399	1698
Western	47	3047	2257	2723
Nilo-Saharan	149	6642	4563	5790
CentralSudanic	44	5114	2590	4079
East	27	3715	2083	3046
EasternSudanic	68	5988	3667	5036
Eastern	14	5103	2098	3871
Nilotic	47	4508	3152	3952
Western	6	5601	1586	3955
Kadugli-Krongo	11	1221	1641	1393
Komuz	9	5209	1656	3752
Koman	6	2542	1411	2078
Saharan	4	3941	1409	2903
Western	3	3553	1322	2638
Songhai	8	1333	1377	1351
Northern	2	807	859	828
Southern	6	580	1220	842

Table B.2: Dates for language groups of Africa

Language group	NOL	ASJP date	3-grams date	CD
Altaic	79	5954	3236	4840
Mongolic	8	2267	1663	2019
Eastern	7	2145	1562	1906
Tungusic	20	1319	2004	1600
Northern	9	1092	1416	1225
Southern	11	1595	1686	1632
Turkic	51	3404	2430	3005
Andamanese	10	4510	1720	3366
GreatAndamanese	8	2122	1493	1864
SouthAndamanese	2	1186	997	1109
Austro-Asiatic	116	3635	3694	3659
Mon-Khmer	97	3406	3481	3437
Aslian	9	2080	1606	1886

EasternMon-Khmer	41	2479	2372	2435
Nicobar	3	3158	1223	2365
NorthernMon-Khmer	29	3259	2271	2854
Palyu	2	2861	501	1893
Viet-Muong	8	2289	1198	1842
Munda	19	2574	1701	2216
NorthMunda	15	1209	1180	1197
SouthMunda	4	2510	1353	2036
Chukotko-Kamchatkan	5	3368	1781	2717
Northern	3	1192	1471	1306
Dravidian	23	2055	2196	2113
Central	3	695	851	759
Northern	3	2030	994	1605
South-Central	7	2447	1501	2059
Southern	10	1894	1628	1785
Hmong-Mien	14	4243	1420	3086
Hmongic	9	2777	1132	2103
Indo-European	218	4348	4855	4556
Baltic	2			
Eastern	2	1469	1169	1346
Celtic	5			
Insular	5	3876	1547	2921
Germanic	30	1745	2417	2021
North	7	1569	1507	1544
West	23	1398	2110	1690
Indo-Iranian	147	3665	3657	3662
Indo-Aryan	93	1996	3076	2439
Iranian	54	2856	2494	2708
Italic	14			
Romance	14	1759	2136	1914
Slavic	16	1157	2092	1540
East	4	1288	1447	1353
South	6	691	1285	935
West	6	820	1413	1063
Japonic	7	1564	1242	1432
Kartvelian	4	2999	1442	2361
Zan	2	596	1042	779
NorthCaucasian	37	7709	3065	5805
EastCaucasian	32	3907	2863	3479
WestCaucasian	5	3649	1245	2663

Sino-Tibetan	165	5261	4445	4926
Chinese	7	2982	1489	2370
Tibeto-Burman	158	4203	4325	4253
Bai	18	1494	717	1175
Himalayish	54	3182	2944	3084
Karen	10	2345	1148	1854
Kuki-Chin-Naga	18	3411	2122	2883
Lolo-Burmese	9	3436	1471	2630
Nungish	3	1955	675	1430
Tangut-Qiang	3	4660	972	3148
Tai-Kadai	68	3252	2009	2742
Hlai	3	2353	726	1686
Kadai	9	2613	981	1944
Kam-Tai	56	2376	1767	2126
Uralic	24	3178	2666	2968
Finnic	6	876	1278	1041
Mordvin	2	800	1015	888
Permian	3	953	891	928
Sami	6	1532	1564	1545
Samoyed	2	2850	1006	2094
Yeniseian	6	2661	1592	2223
AP	2	2762	1172	2110
KA	2	781	981	863
Yukaghir	2	2027	1162	1672

Table B.3: Dates for language groups of Eurasia

Language group	NOL	ASJP date	3-gram date	CD
Amto-Musan	3	2189	997	1700
Arai-Kwomtari	9	7386	2030	5190
Arai(LeftMay)	4	2974	1358	2311
Kwomtari	5	5968	1686	4212
Australian	192	5296	4534	4984
Bunaban	2	1538	1021	1326
Daly	17	3941	1783	3056
Bringen-Wagaydy	10	2320	1344	1920
Malagmalag	4	1635	1169	1444
Murrinh-Patha	3	2747	1074	2061
Djeragan	2	2750	1240	2131
Giimbiyu	3	415	1130	708

Gunwingguan	25	4517	2714	3778
Burarran	3	3612	1442	2722
Enindhilyagwa	3	4746	1331	3346
Gunwinggic	6	2951	1392	2312
Maran	3	2661	1397	2143
Rembargic	2	1925	1030	1558
Yangmanic	2	1609	1240	1458
Pama-Nyungan	122	4295	3958	4157
Arandic	5	1892	1403	1692
Dyirbalic	4	2137	1369	1822
Galgadungic	2	2366	1063	1832
Karnic	6	2851	1610	2342
Maric	9	929	1290	1077
Paman	21	4918	2403	3887
South-West	23	3103	2453	2837
Waka-Kabic	4	2270	1187	1826
Wiradhuric	3	1129	1193	1155
Worimi	2	2473	1237	1966
Yidinic	2	1237	1015	1146
Yuin	3	1503	1306	1422
Yuulngu	16	1555	1991	1734
WestBarkly	3	2631	1442	2144
Wororan	6	2183	1599	1944
Yiwaidjan	4	2882	1401	2275
Yiwaidjic	2	1407	1066	1267
Austronesian	974	3633	6455	4790
Atayalic	2	2664	1269	2092
EastFormosan	4	2392	1489	2022
Malayo-Polynesian	954	3024	6334	4381
Celebic	61	1796	2565	2111
Eastern	44	1710	2120	1878
Kaili-Pamona	3	1076	1033	1058
Tomini-Tolitoli	12	1468	1705	1565
Central-Eastern	581	3111	5655	4154
CentralMalayo-Polynesian	108	2415	3338	2793
EasternMalayo-Polynesian	472	3803	5426	4468
GreaterBarito	57	2031	2450	2203
East	26	1881	1832	1861
Sama-Bajaw	22	1489	1556	1516

West	8	1087	1428	1227
Javanese	3	566	1030	756
Lampung	24	785	1679	1152
LandDayak	3	1510	1151	1363
Malayo-Sumbawan	34	1845	2445	2091
NorthandEast	32	1898	2365	2089
NorthBorneo	17	2016	2047	2029
Melanau-Kajang	2	1372	946	1197
NorthSarawakan	10	2172	1755	2001
Sabahan	5	1333	1269	1307
NorthwestSumatra-BarrierIslands	4	1822	1193	1564
Philippine	151	1830	3463	2500
Bashiic	10	717	1473	1027
Bilic	8	1633	1397	1536
CentralLuzon	3	1252	1042	1166
GreaterCentralPhilippine	75	1326	2718	1897
Minahasan	5	604	1077	798
NorthernLuzon	42	1621	2337	1915
Sangiric	6	484	1100	737
SouthSulawesi	12	970	1545	1206
Bugis	3	884	1074	962
Makassar	5	558	1140	797
Northern	4	345	1057	637
NorthwestFormosan	2	2204	1220	1801
Tsouic	3	2291	1287	1879
WesternPlains	4	2586	1767	2250
CentralWesternPlains	3	2431	1688	2126
Border	16	3453	2201	2940
Taikat	8	2404	1681	2108
Waris	8	2261	1735	2045
CentralSolomons	5	3677	1403	2745
EastBirdsHead-Sentani	13	6615	2047	4742
EastBirdsHead	3	3590	1299	2651
Sentani	9	4101	1606	3078
EastGeelvinkBay	4	3979	1220	2848
EasternTrans-Fly	39	3257	2359	2889
Kaure	2			
KaureProper	2	2665	1180	2056
LakesPlain	26	5279	2230	4029
Rasawa-Saponi	2	3037	1003	2203

Tariku	22	3541	1999	2909
LeftMay	3	2665	1039	1998
Mairasi	4	1196	1287	1233
Nimboran	5	2059	1220	1715
NorthBougainville	2	2925	1175	2208
Pauwasi	7	4102	1794	3156
Eastern	3	2842	1453	2273
Western	4	1774	1271	1568
Piawi	7	3203	1564	2531
Ramu-LowerSepik	20	6942	2500	5121
LowerSepik	9	3411	2032	2846
Ramu	9	4000	1757	3080
Sepik	28	4827	2693	3952
Ndu	9	1227	1242	1233
Nukuma	2	1791	1105	1510
Ram	2	1791	1006	1469
SepikHill	10	3538	1934	2880
Sko	14	4478	1628	3310
Krisa	8	2400	1315	1955
Vanimo	6	1798	1071	1500
SouthBougainville	3	3054	1273	2324
Buin	2	1744	1135	1494
South-CentralPapuan	20	6232	2326	4631
Morehead-UpperMaro	7	5353	1688	3850
Pahoturi	6	2044	1493	1818
Yelmek-Maklew	4	1468	1074	1306
Tor-Kwerba	14	4435	2106	3480
GreaterKwerba	9	4109	1651	3101
Kwerba	6	3852	1394	2844
Orya-Tor	5	3693	1555	2816
Torricelli	26	5754	2876	4574
Kombio-Arapesh	8	3356	1821	2727
Marienberg	9	3339	1991	2786
Monumbo	2	1867	939	1487
Wapei-Palei	5	5386	1612	3839
Trans-NewGuinea	412	6609	5538	6170
Angan	2			
NuclearAngan	2	4523	1021	3087
Asmat-Kamoro	8	2189	1445	1884

Asmat	4	1033	1074	1050
Sabakor	2	567	891	700
Binanderean	5			
Binandere	5	1842	1366	1647
Bosavi	15	2349	1865	2151
Chimbu-Wahgi	10	3470	1701	2745
Chimbu	5	1635	1266	1484
Hagen	3	1505	926	1268
Jimi	2	912	959	931
Duna-Bogaya	2	3004	968	2169
EastStrickland	7	1401	1297	1358
Eleman	9	4851	1465	3463
NuclearEleman	6	1256	1198	1232
Engan	14	2762	1978	2441
Enga	8	2406	1748	2136
Angal-Kewa	4	1555	1146	1387
Finisterre-Huon	19	4136	2308	3387
Finisterre	5	2868	1428	2278
Huon	14	3044	1995	2614
Gogodala-Suki	8	2827	1440	2258
Gogodala	7	1494	1326	1425
InlandGulf	3	2867	1124	2152
Minanibai	2	2197	981	1698
Kainantu-Goroka	24	4847	2608	3929
Gorokan	14	3186	2248	2801
Kainantu	10	3105	1786	2564
Kayagar	4	1285	1063	1194
Kiwaian	14	1436	1789	1581
Kolopom	3	2892	1113	2163
Madang	101	4573	3852	4277
Croisilles	55	4107	3113	3699
RaiCoast	30	3511	2640	3154
SouthAdelbertRange	15	4165	2197	3358
Marind	14	4014	1848	3126
Boazi	8	1597	1315	1481
Yaqay	3	2069	1086	1666
Mek	4	1309	1294	1303
Eastern	3	1425	1177	1323
Mombum	2	1313	1006	1187

Ok-Awyu	21	4272	2263	3448
Awyu-Dumut	9	2916	1641	2393
Ok	12	2534	1796	2231
SoutheastPapuan	25	5286	2235	4035
Goilalan	2	4233	1119	2956
Koiarian	7	2691	1369	2149
Kwalean	6	3032	1218	2288
Mailuan	3	1238	1042	1158
Manubaran	6	1065	1185	1114
Teberan	2	2322	898	1738
Turama-Kikorian	4	3028	1235	2293
Turama-Omatian	3	1580	1122	1392
West	59	5082	3158	4293
Dani	9	1782	1632	1721
EastTimor	3	1916	1080	1573
WestBomberai	3	3497	1200	2555
WestTimor-Alor-Pantar	40	3531	2665	3176
WisselLakes	3	2060	1091	1663
WestPapuan	33	9083	2408	6346
NorthHalmahera	18	2962	1770	2473
Yele-WestNewBritain	2	6293	1097	4163

Table B.4: Dates for language groups of Pacific

Language group	NOL	ASJP date	3-gram date	CD
Algic	27	5554	3183	4582
Algonquian	25	3343	3059	3227
Central	14	2678	2357	2546
Eastern	8	3026	2216	2694
Plains	2	5002	1151	3423
Caddoan	4	4828	1473	3452
Northern	3	3035	1278	2315
Chumash	5	1792	1426	1642
Eskimo-Aleut	9	5084	1895	3777
Eskimo	8	1842	1816	1831
Gulf	3	7859	1102	5089
Hokan	25	4915	2620	3974
Esselen-Yuman	11			
Yuman	11	1865	1672	1786
Northern	13	5666	2095	4202

Karok-Shasta	5	5246	1748	3812
Pomo	7	1226	1042	1151
Iroquoian	7	4855	1998	3684
NorthernIroquoian	6	3176	1886	2647
FiveNations	5	1673	1672	1673
KiowaTanoan	3	3434	1006	2439
Mayan	76	2220	2738	2432
Cholan-Tzeltalan	9	1432	1386	1413
Cholan	5	1148	1122	1137
Tzeltalan	4	511	1006	714
Huastecan	2	1257	946	1129
Kanjobalan-Chujean	8	1225	1326	1266
Chujean	3	1058	965	1020
Kanjobalan	5	803	1030	896
Quichean-Mamean	52	1649	2135	1848
GreaterMamean	29	1492	1729	1589
GreaterQuichean	23	981	1537	1209
Yucatecan	5	790	1071	905
Mopan-Itza	3	887	959	917
Yucatec-Lacandon	2	601	743	659
Misumalpan	3	2774	1009	2050
Mixe-Zoque	14	1407	1551	1466
Mixe	7	900	1193	1020
Zoque	7	787	1208	960
Muskogean	6	1720	1479	1621
Eastern	4	1188	1285	1228
Western	2	345	981	606
Na-Dene	23			
NuclearNa-Dene	22	8532	2145	5913
Athapaskan-Eyak	21	4203	2073	3330
Athapaskan	20	2062	1956	2019
Oto-Manguean	74	6591	3655	5387
Chiapanec-Mangue	2	2445	1195	1933
Chinantecan	4	1935	1063	1577
Mixtecan	9	4542	1471	3283
Mixtec-Cuicatec	7	3140	1313	2391
Trique	2	1024	801	933
Otopamean	7	3654	1555	2793

Otomian	5	2214	1373	1869
Popolocan	17	3036	1900	2570
Chocho-Popolocan	5	2209	1195	1793
Mazatecan	11	775	1522	1081
Subtiaba-Tlapanecan	6	948	1306	1095
Zapotecan	28	3149	2313	2806
Chatino	3	997	922	966
Zapotec	25	1676	2209	1895
Penutian	25	5522	2833	4420
Maiduan	4	1219	1100	1170
OregonPenutian	4	11886	1510	7632
CoastOregon	3	4902	1399	3466
PlateauPenutian	3	4147	1353	3001
Sahaptin	2	2725	1185	2094
Yok-Utian	11	4413	1943	3400
Utian	9	3663	1805	2901
Miwokan	7	2141	1564	1904
Salishan	20	3827	3041	3505
CentralSalish	10	2459	2131	2325
InteriorSalish	6	2980	1978	2569
Siouan	16	6178	2381	4621
SiouanProper	15	3169	2330	2825
Tequistlatecan	2	1212	997	1124
Totonacan	14	1435	1648	1522
Tepehua	3	506	1237	806
Totonac	11	546	1355	878
Uto-Aztecan	82	4018	3167	3669
NorthernUto-Aztecan	11	2576	1934	2313
Numic	7	1737	1570	1669
SouthernUto-Aztecan	71	3472	2831	3209
Aztecan	58			
GeneralAztec	58	1509	2410	1878
Sonoran	13	2400	1869	2182
Wakashan	5	2781	1377	2205
Northern	2	606	717	652
Southern	3	1154	1225	1183
Yuki	2	2500	1000	1885

Table B.5: Dates for language groups of North and Middle America

Language group	NOL	ASJP date	3-gram date	CD
Arauan	7	1764	1497	1655
Arawakan	49			
Maipuran	49	4134	3460	3858
Aymaran	3	1057	1151	1096
Barbacoan	5	3080	1364	2376
Cayapa-Colorado	2	1419	946	1225
Coconucan	2	419	895	614
Cahuapanan	2	1185	1051	1130
Carib	18	2362	2342	2354
Northern	12	2371	1922	2187
Southern	6	2422	1689	2121
Chapacura-Wanham	2	1931	926	1519
Chibchan	22	4400	2741	3720
Aruak	4	2800	1447	2245
Guaymi	3	3286	1012	2354
Kuna	2	820	1036	909
Rama	2	5117	1124	3480
Talamanca	5	2731	1440	2202
Choco	8	2258	1392	1903
Embera	7	875	1313	1055
Chon	2	2774	1108	2091
Guahiban	5	1291	1537	1392
Jivaroan	4	678	1180	884
Katukinan	3	1965	1074	1600
Macro-Ge	26	7266	2864	5461
Ge-Kaingang	13	4989	1947	3742
Yabuti	2	1607	919	1325
Maku	8	3124	1465	2444
Mascoian	3	1718	1499	1628
Mataco-Guaicuru	10	4701	2110	3639
Guaicuruan	5	2909	1536	2346
Mataco	5	2404	1608	2078
Nambiquaran	3	2807	1235	2162
Panoan	19	1853	2268	2023
North-Central	4	2134	1360	1817
Northern	3	1099	1083	1092
South-Central	6	1853	1532	1721
Southeastern	3	920	1051	974

Quechuan	19	1717	1579	1660
QuechuaII	18	974	1440	1165
Tacanan	4	1590	1203	1431
Araona-Tacana	3	1266	1068	1185
Tucanoan	19	2699	2345	2554
EasternTucanoan	13	1241	1801	1471
WesternTucanoan	5	2156	1597	1927
Tupi	47	3585	3004	3347
Monde	5	1712	1262	1528
Munduruku	2	1480	891	1239
Tupari	3	1850	1033	1515
Tupi-Guarani	32	1550	2492	1936
Yuruna	2	951	836	904
Uru-Chipaya	3	1520	1111	1352
Witotoan	7	5491	1813	3983
Boran	3	2271	1362	1898
Witoto	4	2903	1311	2250
Yanomam	8	1319	1547	1412
Zamucoan	3	2765	1304	2166
Zaparoan	3	3178	1399	2449

Table B.6: Dates for language groups of South America

C Appendix to evaluation of string similarity measures

This chapter contains the supplementary information to the tables referred in publication VI.

C.1 Results for string similarity

The tables C.1, C.2, and C.3 show the family averages of Goodman-Kruskal's Gamma, distinctiveness score, and WALS **r** for different string similarity measures.

Family	LDND	LCSD	LDN	LCS	PREFIXD	PREFIX	JCDD	DICED	DICE	JCD
WF										
Tor	0.7638	0.734	0.7148	0.7177	0.7795	0.7458	0.7233	0.7193	0.7126	0.7216
Chi	0.7538	0.7387	0.7748	0.7508	0.6396	0.7057	0.7057	0.7057	0.7057	0.7477
HM	0.6131	0.6207	0.5799	0.5505	0.5359	0.5186	0.4576	0.429	0.4617	0.4384
Hok	0.5608	0.5763	0.5622	0.5378	0.5181	0.4922	0.5871	0.5712	0.5744	0.5782
Tot	1	1	1	1	0.9848	0.9899	0.9848	0.9899	0.9949	0.9848
Aus	0.4239	0.4003	0.4595	0.4619	0.4125	0.4668	0.4356	0.4232	0.398	0.4125
WP	0.7204	0.7274	0.7463	0.7467	0.6492	0.6643	0.6902	0.6946	0.7091	0.697
MUM	0.7003	0.6158	0.7493	0.7057	0.7302	0.6975	0.5477	0.5777	0.6594	0.6213
Sko	0.7708	0.816	0.7396	0.809	0.7847	0.7882	0.6632	0.6944	0.6458	0.6181
ST	0.6223	0.6274	0.6042	0.5991	0.5945	0.5789	0.5214	0.5213	0.5283	0.5114
Sio	0.8549	0.8221	0.81	0.7772	0.8359	0.8256	0.772	0.7599	0.7444	0.7668
Pan	0.3083	0.3167	0.2722	0.2639	0.275	0.2444	0.2361	0.2694	0.2611	0.2306
AuA	0.5625	0.5338	0.5875	0.548	0.476	0.4933	0.5311	0.5198	0.5054	0.5299
Mar	0.9553	0.9479	0.9337	0.9017	0.9256	0.9385	0.924	0.918	0.9024	0.9106
Kad										
May	0.7883	0.7895	0.7813	0.7859	0.7402	0.7245	0.8131	0.8039	0.7988	0.8121
NC	0.4193	0.4048	0.3856	0.3964	0.2929	0.2529	0.3612	0.3639	0.2875	0.2755
Kiw										
Hui	0.9435	0.9464	0.9435	0.9464	0.9464	0.9435	0.8958	0.9107	0.9137	0.8988
LSR	0.7984	0.7447	0.7234	0.6596	0.7144	0.692	0.7626	0.748	0.6484	0.6775
TK	0.7757	0.7698	0.7194	0.7158	0.7782	0.7239	0.6987	0.6991	0.6537	0.6705
LP	0.6878	0.6893	0.7237	0.7252	0.6746	0.7065	0.627	0.6594	0.6513	0.6235
Que	0.737	0.7319	0.758	0.7523	0.742	0.7535	0.7334	0.7335	0.7502	0.7347
NS	0.5264	0.4642	0.4859	0.4532	0.4365	0.3673	0.5216	0.5235	0.4882	0.4968
AA	0.6272	0.6053	0.517	0.459	0.6134	0.5254	0.5257	0.5175	0.4026	0.5162
Ura	0.598	0.5943	0.6763	0.6763	0.5392	0.6495	0.7155	0.479	0.6843	0.7003

MGe	0.6566	0.6659	0.6944	0.716	0.6011	0.662	0.7245	0.7099	0.7508	0.6983
Car	0.325	0.3092	0.3205	0.3108	0.2697	0.2677	0.313	0.3118	0.2952	0.316
Bor	0.7891	0.8027	0.7823	0.7914	0.7755	0.7619	0.7846	0.8005	0.7914	0.7823
Bos										
EA	0.844	0.8532	0.8349	0.8349	0.8716	0.8899	0.8716	0.8716	0.8899	0.8899
TNG	0.6684	0.6692	0.6433	0.6403	0.643	0.6177	0.5977	0.5946	0.5925	0.5972
Dra	0.6431	0.6175	0.6434	0.6288	0.6786	0.6688	0.6181	0.6351	0.655	0.6112
IE	0.7391	0.7199	0.7135	0.6915	0.737	0.7295	0.5619	0.5823	0.6255	0.5248
OM	0.9863	0.989	0.9755	0.9725	0.9527	0.9513	0.9459	0.9472	0.9403	0.9406
Tuc	0.6335	0.623	0.6187	0.6089	0.6189	0.6153	0.5937	0.5983	0.5917	0.5919
Arw	0.5079	0.4825	0.4876	0.4749	0.4475	0.4472	0.4739	0.4773	0.4565	0.4727
NDa	0.9458	0.9578	0.9415	0.9407	0.9094	0.9121	0.8071	0.8246	0.8304	0.8009
Alg	0.5301	0.5246	0.5543	0.5641	0.4883	0.5147	0.4677	0.4762	0.5169	0.5106
Sep	0.8958	0.8731	0.9366	0.9388	0.8852	0.9048	0.8535	0.8724	0.892	0.8701
NDe	0.7252	0.7086	0.7131	0.7017	0.7002	0.6828	0.6654	0.6737	0.6715	0.6639
Pen	0.8011	0.7851	0.8402	0.831	0.8092	0.8092	0.7115	0.7218	0.7667	0.7437
An	0.2692	0.2754	0.214	0.1953	0.2373	0.1764	0.207	0.2106	0.1469	0.2036
Tup	0.9113	0.9118	0.9116	0.9114	0.8884	0.8921	0.9129	0.9127	0.9123	0.9119
Kho	0.8558	0.8502	0.8071	0.7903	0.8801	0.8333	0.8052	0.8146	0.736	0.7378
Alt	0.8384	0.8366	0.85	0.8473	0.8354	0.8484	0.8183	0.8255	0.8308	0.8164
UA	0.8018	0.818	0.7865	0.8002	0.7816	0.7691	0.8292	0.8223	0.8119	0.8197
Sal	0.8788	0.8664	0.8628	0.8336	0.8793	0.8708	0.7941	0.798	0.7865	0.7843
MZ	0.7548	0.7692	0.7476	0.7524	0.7356	0.7212	0.6707	0.6779	0.6731	0.6683

Table C.1: GE for families and measures above average.

Family	JCDD	JCD	TRIGRAMD	DICED	IDENTD	PREFIXD	LDND	LCSD	LDN
Bos	15.0643	14.436	7.5983	10.9145	14.4357	10.391	8.6767	8.2226	4.8419
NDe	19.8309	19.2611	8.0567	13.1777	9.5648	9.6538	10.1522	9.364	5.2419
NC	1.7703	1.6102	0.6324	1.1998	0.5368	1.0685	1.3978	1.3064	0.5132
Pan	24.7828	22.4921	18.5575	17.2441	12.2144	13.7351	12.7579	11.4257	6.8728
Hok	10.2645	9.826	3.6634	7.3298	4.0392	3.6563	4.84	4.6638	2.7096
Chi	4.165	4.0759	0.9642	2.8152	1.6258	2.8052	2.7234	2.5116	1.7753
Tup	15.492	14.4571	9.2908	10.4479	6.6263	8.0475	8.569	7.8533	4.4553
WP	8.1028	7.6086	6.9894	5.5301	7.0905	4.0984	4.2265	3.9029	2.4883
AuA	7.3013	6.7514	3.0446	4.5166	3.4781	4.1228	4.7953	4.3497	2.648
An	7.667	7.2367	4.7296	5.3313	2.5288	4.3066	4.6268	4.3107	2.4143
Que	62.227	53.7259	33.479	29.7032	27.1896	25.9791	23.7586	21.7254	10.8472
Kho	6.4615	6.7371	3.3425	4.4202	4.0611	3.96	3.8014	3.3776	2.1531
Dra	18.5943	17.2609	11.6611	12.4115	7.3739	10.2461	9.8216	8.595	4.8771
Aus	2.8967	3.7314	1.5668	2.0659	0.7709	1.8204	1.635	1.5775	1.4495
Tuc	25.9289	24.232	14.0369	16.8078	11.6435	12.5345	12.0163	11.0698	5.8166
Ura	6.5405	6.1048	0.2392	1.6473	-0.0108	3.4905	3.5156	3.1847	2.1715
Arw	6.1898	6.0316	4.0542	4.4878	1.7509	2.9965	3.5505	3.3439	2.1828
May	40.1516	37.7678	17.3924	22.8213	17.5961	14.4431	15.37	13.4738	7.6795
LP	7.5669	7.6686	3.0591	5.3684	5.108	4.8677	4.3565	4.2503	2.8572
OM	4.635	4.5088	2.8218	3.3448	2.437	2.6701	2.7328	2.4757	1.3643
Car	15.4411	14.6063	9.7376	10.6387	5.1435	7.7896	9.1164	8.2592	5.0205
TNG	1.073	1.216	0.4854	0.8259	0.5177	0.8292	0.8225	0.8258	0.4629
MZ	43.3479	40.0136	37.9344	30.3553	36.874	20.4933	18.2746	16.0774	9.661
Bor	9.6352	9.5691	5.011	6.5316	4.1559	6.5507	6.3216	5.9014	3.8474
Pen	5.4103	5.252	3.6884	3.8325	2.3022	3.2193	3.1645	2.8137	1.5862

MGe	4.2719	4.0058	1.0069	2.5482	1.6691	2.0545	2.4147	2.3168	1.1219
ST	4.1094	3.8635	0.9103	2.7825	2.173	2.7807	2.8974	2.7502	1.3482
Tor	3.2466	3.1546	2.2187	2.3101	1.7462	2.1128	2.0321	1.9072	1.0739
TK	15.0085	13.4365	5.331	7.7664	7.5326	8.1249	7.6679	6.9855	2.8723
IE	7.3831	6.7064	1.6767	2.8031	1.6917	4.1028	4.0256	3.6679	1.4322
Alg	6.8582	6.737	4.5117	5.2475	1.2071	4.5916	5.2534	4.5017	2.775
NS	2.4402	2.3163	1.1485	1.6505	1.1456	1.321	1.3681	1.3392	0.6085
Sko	6.7676	6.3721	2.5992	4.6468	4.7931	5.182	4.7014	4.5975	2.5371
AA	1.8054	1.6807	0.7924	1.2557	0.4923	1.37	1.3757	1.3883	0.6411
LSR	4.0791	4.3844	2.2048	2.641	1.5778	2.1808	2.1713	2.0826	1.6308
Mar	10.9265	10.0795	8.5836	7.1801	6.4301	5.0488	4.7739	4.5115	2.8612
Alt	18.929	17.9969	6.182	9.1747	7.2628	9.4017	8.8272	7.9513	4.1239
Sep	6.875	6.5934	2.8591	4.5782	4.6793	4.3683	4.1124	3.8471	2.0261
Hui	21.0961	19.8025	18.4869	14.7131	16.1439	12.4005	10.2317	9.2171	4.9648
NDa	7.6449	7.3732	3.2895	4.8035	2.7922	5.7799	5.1604	4.8233	2.3671
Sio	13.8571	12.8415	4.2685	9.444	7.3326	7.8548	7.9906	7.1145	4.0156
Kad	42.0614	40.0526	27.8429	25.6201	21.678	17.0677	17.5982	15.9751	9.426
MUM	7.9936	7.8812	6.1084	4.7539	4.7774	3.8622	3.4663	3.4324	2.1726
WF	22.211	20.5567	27.2757	15.8329	22.4019	12.516	11.2823	10.4454	5.665
Sal	13.1512	12.2212	11.3222	9.7777	5.2612	7.4423	7.5338	6.7944	3.4597
Kiw	43.2272	39.5467	46.018	30.1911	46.9148	20.2353	18.8007	17.3091	10.3285
UA	21.6334	19.6366	10.4644	11.6944	4.363	9.6858	9.4791	8.9058	4.9122
Tot	60.4364	51.2138	39.4131	33.0995	26.7875	23.5405	22.6512	21.3586	11.7915
HM	8.782	8.5212	1.6133	4.9056	4.0467	5.7944	5.3761	4.9898	2.8084
EA	27.1726	25.2088	24.2372	18.8923	14.1948	14.2023	13.7316	12.1348	6.8154
Average	15.0501	13.9673	9.448	9.4416	8.163	7.5359	7.3189	6.7042	3.7943

Table C.2: Dist for families and measures above average

Family	LDND	LCSD	LDN	LCS	PREFIXD	PREFIX	DICED	DICE	JCD	JCDD	TRID
NDe Bos	0.5761	0.5963	0.5556	0.5804	0.5006	0.4749	0.4417	0.4372	0.4089	0.412	0.2841
NC Hok Pan	0.4569	0.4437	0.4545	0.4398	0.3384	0.3349	0.3833	0.3893	0.3538	0.3485	0.2925
Chi	0.5735	0.5775	0.555	0.5464	0.5659	0.5395	0.5616	0.5253	0.5593	0.5551	0.4752
Tup	0.7486	0.7462	0.7698	0.7608	0.6951	0.705	0.7381	0.7386	0.7136	0.7125	0.6818
WP	0.6317	0.6263	0.642	0.6291	0.5583	0.5543	0.5536	0.5535	0.5199	0.5198	0.5076
AuA Que	0.6385	0.6413	0.5763	0.5759	0.6056	0.538	0.5816	0.5176	0.5734	0.5732	0.5147
An	0.1799	0.1869	0.1198	0.1003	0.1643	0.0996	0.1432	0.0842	0.1423	0.1492	0.1094
Kho	0.7333	0.7335	0.732	0.7327	0.6826	0.6821	0.6138	0.6176	0.5858	0.582	0.4757
Dra	0.5548	0.5448	0.589	0.5831	0.5699	0.6006	0.5585	0.589	0.5462	0.5457	0.5206
Aus Tuc	0.2971	0.2718	0.3092	0.3023	0.2926	0.3063	0.2867	0.257	0.2618	0.2672	0.2487
Ura Arw May	0.4442	0.4356	0.6275	0.6184	0.4116	0.6104	0.2806	0.539	0.399	0.3951	0.1021
LP	0.41	0.4279	0.4492	0.4748	0.3864	0.4184	0.3323	0.336	0.3157	0.3093	0.1848
OM Car MZ	0.8095	0.817	0.7996	0.7988	0.7857	0.7852	0.7261	0.7282	0.6941	0.6921	0.6033

TNG	0.5264	0.5325	0.4633	0.4518	0.5	0.472	0.469	0.4579	0.4434	0.4493	0.3295
Bor											
Pen	0.8747	0.8609	0.8662	0.8466	0.8549	0.8505	0.8531	0.8536	0.8321	0.8308	0.7625
MGe	0.6833	0.6976	0.6886	0.6874	0.6086	0.6346	0.6187	0.6449	0.6054	0.6052	0.4518
ST	0.5647	0.5596	0.5435	0.5261	0.5558	0.5412	0.4896	0.4878	0.4788	0.478	0.3116
IE	0.6996	0.6961	0.6462	0.6392	0.6917	0.6363	0.557	0.5294	0.5259	0.5285	0.4541
TK	0.588	0.58	0.5004	0.4959	0.5777	0.4948	0.5366	0.4302	0.5341	0.535	0.4942
Tor	0.4688	0.4699	0.4818	0.483	0.4515	0.4602	0.4071	0.4127	0.375	0.3704	0.3153
Alg	0.3663	0.3459	0.4193	0.4385	0.3456	0.3715	0.2965	0.3328	0.291	0.2626	0.1986
NS	0.6118	0.6072	0.5728	0.5803	0.5587	0.5118	0.578	0.5434	0.5466	0.5429	0.4565
Sko	0.8107	0.8075	0.806	0.7999	0.7842	0.7825	0.6798	0.6766	0.6641	0.6664	0.5636
AA	0.6136	0.6001	0.4681	0.431	0.6031	0.4584	0.5148	0.3291	0.4993	0.4986	0.4123
LSR	0.5995	0.5911	0.6179	0.6153	0.5695	0.5749	0.5763	0.5939	0.5653	0.5529	0.5049
Mar	0.654	0.6306	0.6741	0.6547	0.6192	0.6278	0.568	0.5773	0.5433	0.5366	0.4847
Alt	0.8719	0.8644	0.8632	0.8546	0.8634	0.8533	0.7745	0.7608	0.75	0.7503	0.6492
Hui	0.6821	0.68	0.6832	0.6775	0.6519	0.6593	0.5955	0.597	0.5741	0.5726	0.538
Sep	0.6613	0.656	0.6662	0.6603	0.6587	0.6615	0.6241	0.6252	0.6085	0.6079	0.5769
NDa	0.6342	0.6463	0.6215	0.6151	0.6077	0.5937	0.501	0.5067	0.4884	0.4929	0.4312
Sio											
Kad											
WF											
MUM											
Sal	0.6637	0.642	0.6681	0.6463	0.6364	0.6425	0.5423	0.5467	0.5067	0.5031	0.4637
Kiw											
UA	0.9358	0.9332	0.9296	0.9261	0.9211	0.9135	0.9178	0.9148	0.8951	0.8945	0.8831
Tot											
EA	0.6771	0.6605	0.6639	0.6504	0.6211	0.6037	0.5829	0.5899	0.5317	0.5264	0.4566
HM											
Average	0.619	0.6151	0.6126	0.6069	0.5859	0.5784	0.5495	0.5449	0.5322	0.5302	0.4495

Table C.3: RW for families and measures above average. TRID refers to TRIGRAMD.