Taraka Rama

Vocabulary lists in computational historical linguistics

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

Editor: Lars Borin

Språkbanken
Department of Swedish
University of Gothenburg

25 • 2014

Taraka Rama

Vocabulary lists in computational historical linguistics

Gothenburg 2014

Data linguistica 25
ISBN 978-91-87850-52-3
ISSN 0347-948X

Printed in Sweden by
Reprocentralen, Campusservice Lorensberg, University of Gothenburg
2014

Typeset in LaTeX 2_ε by the author

Cover design by Kjell Edgren, Informat.se

Front cover illustration:
A network representation of relations between Dravidian languages.
by Rama and Kolachina (2013) ©

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, we work with vocabulary lists for addressing some classical problems in historical linguistics such as discriminating related languages from unrelated languages, assigning possible dates to splits in a language family, employing structural similarity for language classification, and providing an internal structure to a language family. In this thesis, we compare the internal structure inferred from vocabulary lists to the family tree structure inferred through the comparative method. We also explore the ranking of lexical items in the widely used Swadesh word list and compare our ranking to another quantitative reranking method and short lists composed for discovering long-distance genetic relationships. We also 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 criticism of constant rate of lexical change assumed by glottochronology. An important conclusion from these results is that n-gram approaches can be used for different historical linguistic purposes. The field is undergoing a shift from – the application of computational methods to – short, hand-crafted vocabulary lists to automatically extracted word lists from corpora. Thus, we also experiment with parallel corpora for automatically extracting cognates to infer a family tree from the cognates.

SAMMANFATTNING

Datorbaserad analys av historiska data har gjort stora framsteg under det senaste decenniet. I denna licentiatuppsats använder vi ordlistor för att ta oss an några klassiska problem inom historisk lingvistik. Exempel på sådana problem är hur man avgör vilka språk som är släkt med varandra och vilka som inte är det, hur man tidsbestämmer rekonstruerade urspråk, hur man klassificerar språk på grundval av strukturella likheter och skillnader, samt hur man sluter sig till den interna strukturen i en språkfamilj (dess ‘familjeträd’).

I uppsatsen jämför vi metoder som använts för att postulera språkliga familjeträd. Specifikt jämför vi ordlistebaserade metoder med den traditionella komparativa metoden, som även använder andra språkliga drag för jämförelsen. Med fokus på ordlistor jämför vi den ofta använda Swadesh-listan med alternativa listor föreslagna i litteraturen eller framtagna i vår egen forskning, med avseende på deras användbarhet för att angripa de nämnda historisk-lingvistiska problemen.

Vi visar också i experiment att valet av likhetsmått är mycket betydelsefullt när strängjämförelser används för att bestämma den interna strukturen i en språkfamilj eller för att skilja besläktade och obesläktade språk åt. Ett viktigt resultat av dessa experiment är att n-grambaserade metoder lämpar sig mycket väl för flera olika språkhistoriska ändamål.

De metoder för språklig datering som presenteras här kan användas för att tidsbestämma nya språkfamiljer, dock utan att vara beroende av antagandet att förändringen av ett språks basordförråd är konstant över tid, ett hårt kritiserat antagande som ligger till grund för glottokronologin som den ursprungligen formulerades.

Metodologiskt har man inom området nu börjat utforska möjligheten att övergå från att arbeta med korta, på förhand givna ordlistor till att tillämpa språkteknologiska metoder på stora språkliga material, t ex hela (traditionella) lexikon eller strukturerat språkligt material som extraheras ur flerspråkiga korpusar. I uppsatsen utforskas användningen av parallella korpusar för att automatiskt finna ord med ett gemensamt ursprung (kognater) och därefter härleda ett språkligt familjeträd från kognatlistorna.

ACKNOWLEDGEMENTS

I thank my supervisor, Lars Borin for taking me in at the beginning of my stay in Sweden and also providing me with a place to live. We had interesting discussions on historical linguistics in the meetings. I thank my assistant supervisor, Søren Wichmann for inviting me to Leipzig and for all the interesting discussions on historical linguistics. He always read any piece of work I wrote and helped me improve it. I thank my assistant supervisor, Markus Forsberg for all the support. I thank all my collaborators Eric Holman, Prasanth Kolachina, and Sudheer Kolachina for all the critic on the work done and published during my stay in Sweden. I thank all the colleagues (past and present) at Språkbanken and Department of Swedish for the nice environment. The paper presentations would not have been possible without the generous funding of Graduate School in Language Technology, Centre for Language Technology, and Digital Areal Linguistics project. I thank all the colleagues (past and present) in these places for the environment. I thank Harald Hammarström for comments on my papers. I thank Anju Saxena for all the support. I also thank the colleagues at MPI-EVA, Leipzig and Alpha-Informatica, Groningen for the nice environment. Finally, this work would not have been possible without the support of my family and friends across the world. I thank my family and friends for all the support through my stay in Sweden.

CONTENTS

Appendices

Part I

Introduction to the thesis

1 INTRODUCTION

This licentiate 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 dating of language families, structural similarity vs genetic similarity, and language classification.

There are more than $7,000$ languages in this world (Lewis, Simons and Fennig 2013) and more than $100,000$ unique languoids (Nordhoff and Hammarström 2012; it is known as *Glottolog*) where a languoid is defined as a set of documented and closely related linguistic varieties. 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 descend from a common language? Did language emerge from a single source (*monogenesis*) or from multiple sources at different times (*polygenesis*)? A less ambitious question would be if there are any relations between these languages? Or 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 these questions. These questions come under the scientific discipline of historical linguistics. More specifically, this thesis operates in the subfield of computational historical linguistics.

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

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 languages (Trask 2000: 150). This branch works towards identifying the not-so-apparent relations between languages. The branch has succeeded in identifying the relation between languages spoken in the Indian sub-continent, the Uyghur region of China, and Europe; the languages spoken in Madagascar islands and the remote islands in the Pacific Ocean.

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*. Comparativists 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 lies within the area of computational historical linguistics which relates to the application of computational techniques to address the traditional problems in historical linguistics.

1.1.2 What is computational historical linguistics?

The use of mathematical and statistical techniques to classify languages (Kroeber and Chrétien 1937) and evaluate the language relatedness hypothesis (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 between 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). The cognates were not identified through a computer but by a manual procedure beforehand to arrive at the pair-wise cognate counts.

Hewson 1973 (see Hewson 2010 for a more recent description) can be considered the first such study where computers were used to reconstruct the words of Proto-Algonquian (the common ancestor of Algonquian language family). The 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 project an ancestral form (proto-form; represented by a *) for a word form by searching through all possible sound-correspondences. The projected proto-forms for each language are 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.

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), computational and evolutionary aspects of language (Nowak, Komarova and Niyogi 2002; Niyogi 2006).

In practice, historical linguists work with word lists – selected words which are not nursery forms, onomatopoeic forms, chance similarities, and borrowings (Campbell 2003) – for the majority of the time. 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 the 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 in the Proto-Indo-European (PIE) is a classical case of internal reconstruction. Saussure applied internal reconstruction to explain the aberrations in the reconstructed root structures of PIE.

PIE used vowel alternations such as English *sing/sang/sung* – also known as *ablaut or apophony* – for grammatical purposes (Trask 1996: 256). The general pattern for root structures was CVC with V reconstructed as *e. However there were exceptions to the reconstructed root of the forms such as C$\bar{V}$- or VC- where V could be *a or *o. Saussure conjectured that there were three consonants: h_1, h_2, h_3 in pre-PIE. Imagining each consonant as a function which operates on vowels **e, **a and **o; h_1 would render **e > *e; h_2 renders **e > *a; h_3 renders **e > *o.[2] Finally, the consonant in pre-vocalic position affected the vowel quality and in post-vocalic position, it also affected the preceding vowel length through compensatory lengthening. This conjecture was corroborated through the discovery of the [ḫ] consonant in Hittite texts.

The following excerpt from the *Lord's Prayer* shows the differences between Old English (OE) and current-day English (Hock 1991: 2–3):

Fæder ūre þū þe eart on heofonum,
Sī þīn nama ġehālgod.

'Father of ours, thou who art in heavens,
Be thy name hallowed.'

In the above excerpt, Old English (OE) *eart* is the ancestor to English *art* 'are' which is related to PIE *h_1*er*-. The OE *sī* (related to German *sind*) and English *be* are descendants from different PIE roots *h_1*es*- and *$b^h uh_2$- but serve the same purpose.

The work reported in this thesis attempts to devise and apply computational techniques (developed in LT) to both hand-crafted word lists as well as automatically extracted word lists from corpora.

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 the 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 the 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.[3] 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).

[2] ** denotes a pre-form in the proto-language.

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

The new wave of CHL publications are co-authored by linguists, computer scientists, computational linguists, physicists and evolutionary biologists. Except for sporadic efforts (Kay 1964; Sankoff 1969; Klein, Kuppin and Meives 1969; Durham and Rogers 1969; Smith 1969; Wang 1969; Dobson et al. 1972; Borin 1988; Embleton 1986; 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, based on the lexical cognate data of Indo-European family to infer the family tree. In LT, Bouchard-Côté et al. (2013) employed Bayesian techniques to reconstruct Proto-Austronesian forms for a 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 can be 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. Lexical borrowings can give a 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 related to the Indo-Aryan languages than the other languages of IE family. One task of historical linguists is to identify borrowings between languages which are known to have contact. A much generalization of the task of identifying borrowings between languages with no documented contact history. Chance similarities are called *false friends* by historical linguists. One famous example from Bloomfield 1935 is Modern Greek *mati* and Malay *mata* 'eye'. However, these languages are unrelated and the words are similar only through chance resemblance.

The word pair Swedish *ingefära* and Sanskrit *sr̥ngavera* 'ginger' have similar shape and the same meaning. However, Swedish borrowed the word from a different source and nativized the word to suit its own phonology. It is known that Swedish never had any contact with Sanskrit speakers and still has this word as a cultural borrowing. Another task would be to automatically identify such indirect borrowings between languages with no direct contact (Wang and Minett 2005). Nelson-Sathi et al. (2011) applied a network model to detect the hidden borrowing in the basic vocabulary lists of Indo-European.

The task of automated language identification (Cavnar and Trenkle 1994) can be related to the task of automated language classification. A 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 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 to. 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 input to a clustering algorithm to infer a hierarchical structure to 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),[4] orthographic changes and grammaticalization over time through the analysis of corpora (Borin et al. 2013).

- *Application in SMT (Statistical Machine Translation)*. SMT techniques are applied to annotate historical corpora, Icelandic from the 14th century, through current-day Icelandic (Pettersson, Megyesi and Tiedemann 2013). Kondrak, Marcu and Knight (2003) employ cognates in SMT

[4] How lexical items acquire a different meaning and function over time. Such as Latin *hostis* 'enemy, foreigner, and stranger' from PIE's original meaning of 'stranger'.

models to improve the translation accuracy. Guy (1994) designs an algorithm for identifying cognates in bi-lingual word lists and attempts to apply it in machine translation.

1.2 Questions, answers, and contributions

This thesis aims to address the following problems in historical linguistics through the application of computational techniques from LT and IE/IR:

I. *Corpus-based phylogenetic inference.* In the age of *big data* (Lin and Dyer 2010), can language relationships be inferred from parallel corpora? *Paper I* entitled *Estimating language relationships from a parallel corpus* presents results on inferring language relations from the parallel corpora of the European Parliament's proceedings. We apply three string similarity techniques to sentence-aligned parallel corpora of 11 European languages to infer genetic relations between the 11 languages. The paper is co-authored with Lars Borin and is published in *NODALIDA 2011* (Rama and Borin 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. *Paper 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 SMT, automated language identification, and automated drug detection (Kondrak and Dorr 2006) – to determine the concepts which are resistant to the effects of time and geography. The results suggest that the ranked item list agrees largely with two other vocabulary lists proposed for identifying long-distance relationship. The paper is co-authored with Lars Borin and is accepted for publication 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? *Paper III* titled *How good are typological distances for determining genealogical relationships among languages?* applies 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 *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 associated with glottochronology (a severely 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). *Paper 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. *Comparison of string similarity measures for automated language classification.* A researcher attempting to carry out an automatic language classification 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 internal structure of the related languages? *Paper V*, *Evaluation of similarity measures for automatic language classification* is a book chapter under review for a proposed edited volume. The paper 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 submitted to a edited volume, *Sequences in Language and Text* (Rama and Borin 2014).

The contributions of the thesis are summarized below:

- Paper I should actually be listed as the last paper since it works with automatically extracted word lists – the next step in going beyond handcrafted word lists (Borin 2013a). The experiments conducted in the paper show that parallel corpora can be used to automatically extract cognates (in the sense used in historical linguistics) and then used to infer a phylogenetic tree.

- Paper II develops 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.

- Paper III attempts to address the following three tasks: (a) Comparison of vector similarity measures for computing typological distances; (b) correlating typological distances with genealogical classification derived from historical linguistics; (c) correlating typological distances with the lexical distances computed from 40-word Swadesh lists. The paper also uses graphical devices to show the strength and direction of correlations.

- Paper IV introduces 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.

- It has been noted 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 very well researched topic in computer science (Smyth 2003) and computer scientists developed various string similarity measures for many practical applications. There is certainly a gap in CHL regarding the performance of other string similarity measures in the tasks of automatic language classification and inference of internal structures of language families. Paper V attempts to fill this gap. The paper compares the performance of 14 different string similarity techniques for the aforementioned purpose.

1.3 Overview of the thesis

The thesis is organized as follows. The first part of the thesis gives an introduction to the papers included in the second part of the thesis.

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 statistical LT learning paradigm of semisupervised learning (Yarowsky 1995; Abney 2004, 2010). Subsequently, the chapter proceeds to discuss the related computational work in the domain of automated language classification. We also propose a language classification system which employs string similarity measures for discriminating related languages from unrelated languages and internal classification. Any classification task requires the selection of suitable techniques for evaluating a system.

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

Chapter 4 summarizes and concludes the introduction to the thesis and discusses future work.

Part II of the thesis consists of four peer-reviewed publications and a book chapter under review. Each paper is reproduced in its original form leading to slight repetition. Except for paper II, rest of the papers are presented in the chronological order of their publication. Paper II is placed after paper I since paper II focuses on ranking of lexical items by genetic stability. The ranking of lexical items is an essential task that precedes the CHL tasks presented in papers III–V.

All the experiments in the papers I, II, IV, and V were conducted by me. The experiments in paper III were designed and conducted by myself and Prasanth Kolachina. The paper was written by myself and Prasanth Kolachina. In papers I, II, and V, analysis of the results and the writing of the paper were performed by myself and Lars Borin. The experiments in paper IV were designed and performed by myself. I am the sole author of paper IV.

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

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

2. 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.

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. *QITL* 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.

2 COMPUTATIONAL HISTORICAL LINGUISTICS

This chapter is devoted to an in-depth survey of the terminology used in the papers listed in part II of the thesis. This chapter covers related work in 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 *Ethnologue* (Lewis, Simons and Fennig 2013) falling into more than 400 families (Hammarström 2010). The following questions arise 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 common and differing linguistic aspects in a language family?
- How do we measure and arrive at a numerical estimate of the differences and diversity? What are the units of such comparison?
- How and why do these differences arise?

The above questions can be addressed in the recent frameworks proposed in 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.

The nineteenth century linguist Schleicher (1853) proposed the *stammbaum* (family tree) device to show the differences as well as similarities between languages. 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 Denisova cave in Russia; henceforth referred to 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 population of species which over the time leads to splitting of the same species into different species. The evolutionary change is a two-step process:

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).

- The generation of variation in the replication process.
- Selection of a variant from the pool of variants.

Dawkins (2006) employs the selfish-gene concept that the organism is only a vector for the replication of the gene. The gene itself is generalized as a replicator. Dawkins and Hull differ 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. 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 pressure for mutual understanding and pressure to confirm to a standard variety cause imperfect replication in Croft's model. The speaker himself selects the variants fit for production whereas, Nettle (1999a) argues that functional pressure also operates in the selection of variants.

The iterative mounting differences induced through generations of imperfect replication cause linguistic diversity. Nettle (1999a: 10) lists three different types of linguistic diversity:

- *Language diversity*. This is simply the number of languages present in a given geographical area. New Guinea has the highest geographical diversity with more than 800 languages spoken in a small island whereas Iceland has only one language (not counting the immigration in the 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 53 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 fourth measure of diversity or differences is based on phonology. 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. Rama and Borin (2011) employ phonotactic similarity for the genetic classification of 11 European languages.

Consider the Scandinavian languages Norwegian, Danish and Swedish. All the three languages are mutually intelligible (to a certain degree) yet are called different languages. How different are these languages or how distant are these languages from each other? Can we measure the pair-wise distances between these languages? In fact, Swedish dialects such as Pitemål and Älvdalska are so different from Standard Swedish that they can be counted as different languages (Parkvall 2009).

In an introduction to the volume titled *Approaches to measuring linguistic differences*, Borin (2013b: 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 a much thinner samples of a single language. Namely, they work with language varieties (dialects) spoken in different sites in the geographical area where the language is spoken.[5] For instance, a Swedish speaker from Gothenburg can definitely communicate with a Swedish speaker of Stockholm. However, there are differences between these varieties and a dialectologist works towards charting the dialectal contours of a language.

At a higher level, the three Scandinavian languages are mutually intelligible to a certain degree but are listed as different languages due to political reasons. Consider the inverse case of Hindi, a language spoken in Northern India. The language extends over a large geographical area but the languages spoken in Eastern India (Eastern Hindi) are not mutually intelligible with the languages spoken in Western India (Western Hindi). Nevertheless, these languages are referred to as Hindi (Standard Hindi spoken by a small section of the Northern Indian population) due to political reasons (Masica 1993).

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 the successful 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, its 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

[5] *Doculect* is the term that has become current and refers to a language variant described in a document.

and hence, employed in computing 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 where a sound becomes less consonant like. Consonants can undergo a shift from right to left 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 **flosis* > *floris* genitive form of 'flower'. 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 left to right on each 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 is in a South-Central Dravidian language, Pengo. The word in 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. When a consonant is inserted between two consonants. For instance, Cypriot Arabic developed a [k] as in **pjara* > *pkjara* (Crowley and Bowern 2009: 31).
Epenthesis. When a vowel is inserted into a middle of a word. Tamil inserts 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. Tamil *ulakam* < Sanskrit *lōkam* ‘world’, *aracan* < *rājan* ‘king’ (Krishnamurti 2003: 476).

2.2.1.4 Metathesis

Two sounds swap their position in this change. Proto-Dravidian (PD) did not allow apical consonants such as *ṭ, ṯ, l, ḷ, ẓ, r* in the word-initial position. However, Telugu allows *r, l* in the word-initial position. This exception developed due to the process of metathesis. For instance, PD **iraṇṭu* > *reṇḍu* ‘two’ where the consonant [r] swapped its position with the preceding vowel [i] (Krishnamurti 2003: 157). 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 Vowel breaking

A vowel can change into a diphthong and yields an extra glide which can be before- (on-glide) or off-glide. An example from Dravidian is the Proto-South Dravidian form **oṭay* > Toda *waṛ* 'to break'; **o* > *wa* before *-ay*.

2.2.1.7 Assimilation

In this sound change, a sound becomes more similar to the sound preceding or after it. In some cases, a sound before exactly the same as the sound next to it – *complete assimilation*; otherwise, it copies some of the phonetic features from the next 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'.[6] Palatalization is a type of assimilation where a consonant preceding a front vowel develops palatal feature, such as [k] > [c]. For example, Telugu shows palatalization from PD: Telugu *cēyi* 'hand'< **key* < **kay* (Krishnamurti 2003: 128).

2.2.1.8 Dissimilation

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

2.2.1.9 Some important sound changes

This subsection deals with some identified sound changes from the Indo-European 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 sound change in a subset of South-Central Dravidian languages which is on-going and did not affect many of the lexical items suitable for sound change (Krishnamurti 1978).

[6] This example is given by B. Lakshmi Bai.

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-* ~ θ*reis* ‘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 next important sound change in IE linguistics is the Grassmann’s law. As mentioned above, Grassmann’s law (GL) states that whenever two syllables (within the same root or when reduplicated) are adjacent to each other, with aspirated stops, 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* > $labd^ha$ ‘sieze’, *dah+ta* > $dagd^ha$ ‘burnt’, bud^h*+ta* > $budd^ha$ ‘awakened’ (Trask 2000: 38).

Together, BL and GL received much attention due to their order of application in the Indic languages. One example is the historical derivation of dug^hdas in Sanskrit. The first solution is to posit $*d^hug^h{+}t^has \overset{BL}{\to} *d^hug^hd^has$

$\xrightarrow{GL}$ **dug^{h}d^{h}as* $\xrightarrow{deaspiration}$ *dugd^{h}as*. Reversing the order of BL and GL yields the same output. Collinge (1985: 49–52) summarizes recent efforts to explain all the roots in Indic branch using a particular rule application order of BL and GL. The main take-away from the GL debate is that the reduplication examples show the clearest deaspiration in first syllable. For instance, *d^h – d^h > d – d^h* in Sanskrit *da-d^{h}ā-ti* ‘to set’, reduplicated present. A loss of second syllable aspiration immediately before /s/, /t/ (Beekes 1995: 128). An example of this sound change from Sanskrit is: *dáh-a-ti* ‘burn’ < PIE **d^{h}ag^h-*, but 3 sg. s-aor. *á-d^{h}āk* < **-dhāk-s-t*.

An example of the application of BL and GL is: *budd^{h}a* can be explained as PIE **b^{h}ewd^h* (e-grade) $\xrightarrow{GL}$ Sanskrit *bud^h* (Ø-grade); *bud^h+ta* $\xrightarrow{BL}$ *budd^{h}a* ‘awakened’ (Ringe 2006: 20).

Another well-known sound change in 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*, *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).

2.2.1.10 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 is by Simard, Foster and Isabelle (1993) employ similar words[7] 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.[8] Covington (1996) employs phonetic features for measuring the change between cognates. The rest of the section introduces Levenshtein distance (Levenshtein 1966) and the other orthographic measures for quantifying the similarity between words. I will also make an attempt at explaining the linguistic motivation for using these measures and their limitations.

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. 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. Kessler (2005) gives a summary of various phonetic algorithms developed for the historical comparison of word forms.

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

1. Represent each symbol as a vector of phonetic features.
2. Compare the vectors of phonetic features belonging to the dissimilar symbols using Manhattan distance, Hamming distance or Euclidean distance.

A feature in a feature vector can be represented as a 1/0 bit or a value on a continuous (Kondrak 2002a) or ordinal (Grimes and Agard 1959) scale. An ordinal scale implies an implicit hierarchy in the phonetic features – place of articulation and manner of articulation. Heeringa (2004) uses a binary feature-valued

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

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

system to compare Dutch dialects. Rama and Singh (2009) use the phonetic features of the Devanagari alphabet 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 unnatural and explainable chain of sound changes is the Armenian *erku* (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 2014).

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 input word representation format to vanilla LD. 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 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 through the Grassmann's Law example. Grassmann's law is a case of distant dissimilation which cannot be retrieved by LD.

There are string similarity measures which work at least as well as LD. A few such measures are Dice, Longest common subsequence ratio (Tiedemann 1999), and Jaccard's measure. Dice and Jaccard's index are related measures which can handle a 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, bigrams 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) observes that semantic change also includes 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 under the section 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 in this change. Originally, *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 still survives in expressions such as *sweetmeat* and *One man’s meat is another man’s poison*. Tamil *kili* ‘bird’ ~ Telugu *chili-* ‘parrot’ is another example of narrowing.

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'.

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' ~ Latin *rota* 'wheel'.

2.2.2.2 Lexical change

Languages acquire new words through the mechanism of *borrowing* and *neologisms*. Borrowing is broadly categorized into lexical borrowing (loanwords) and loan translations. Lexical borrowing usually 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 such cases are called *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 names – *volt, ohm, vandalize* (from Vandals); place names – Swedish *persika* 'peach' < Persia; from compounding – *braindead*; from derivation – *boombox*; amalgamation – *altogether, always, however*; from clipping – *gym* < *gymnasium, bike* < *bicycle*, and *nuke* < *nuclear*.

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.[9] 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 'parrot' is an example of a semantic shift which occurred in a daughter language (Telugu) from the Proto-Dravidian's original meaning of 'bird'.

The work of Kondrak 2001, 2004, 2009 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. As-

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

suming 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. This step restricts the comparison of grammatically over-loaded forms and the identification of grammaticalization.
- *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 were few other works to compute semantic distance between languages based on bilingual dictionaries (Cooper 2008; Eger and Sejane 2010).

The major deficiency in Kondrak's work is the restriction on the mobility of meaning across syntactic categories and the restriction to nouns. In contrast, comparative linguists also work with comparing and reconstructing of bound morphemes and their functions. Moreover, grammaticalization is not recognized in this framework. Finally, Kondrak's algorithms require comparative dictionaries as an input, which require a great deal of human effort. This seems to be remedied to a certain extent in the work of Tahmasebi (2013) and Tahmasebi and Risse (under submission).

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 also 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. This is known as the time limit of the comparative method (Harrison 2003) where the comparative method does not work for recovering temporally deep – greater than 8000 years (Nichols 1992) – language change.

The rest of the section describes the ingredients which go into the comparative method, models of language change, examples of how few families were established through the comparative method, and the mechanized parts of the comparative method.

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. The next subsection describes the selection process of vocabulary items and examples of grammatical correspondences for positing 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 – *eṇḍa*, *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 (Ø-grade)	*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 reported 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, IE 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. Later comparison showed that Armenian borrowed heavily from Iranian yielding the earlier conclusion that Armenian is a language within Iranian subfamily. Later grammatical comparison, however, 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 weigh the binary vs. ternary splitting hypotheses at the top-most node.

2.3.2 The comparative method and reconstruction

The previous subsection introduced the three sources for accumulating evidence for consolidating the genetic relation proposals between languages. This section summarizes the working of comparative method and the procedure for reconstructing the proto-language as well as the intermediary proto-languages. The comparative method has been described in various articles 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. The rest of the section summarizes the various steps and the models of language change with illustrations.

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/ ~ Latin /*d*/ in *theos* ~ *deus* (Fox 1995: 66) – we know that Greek /t^h/ should correspond to Latin /*f*/ in word-initial position. These sound correspondences are then used to search for plausible cognates across the languages. Meillet requires that a plausible cognate should occur in at least three languages to label the cognate set as plausible. In the next step, a possible proto-phoneme for a sound correspondences set is posited. For 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 is *g^{wh}. This reconstruction comes

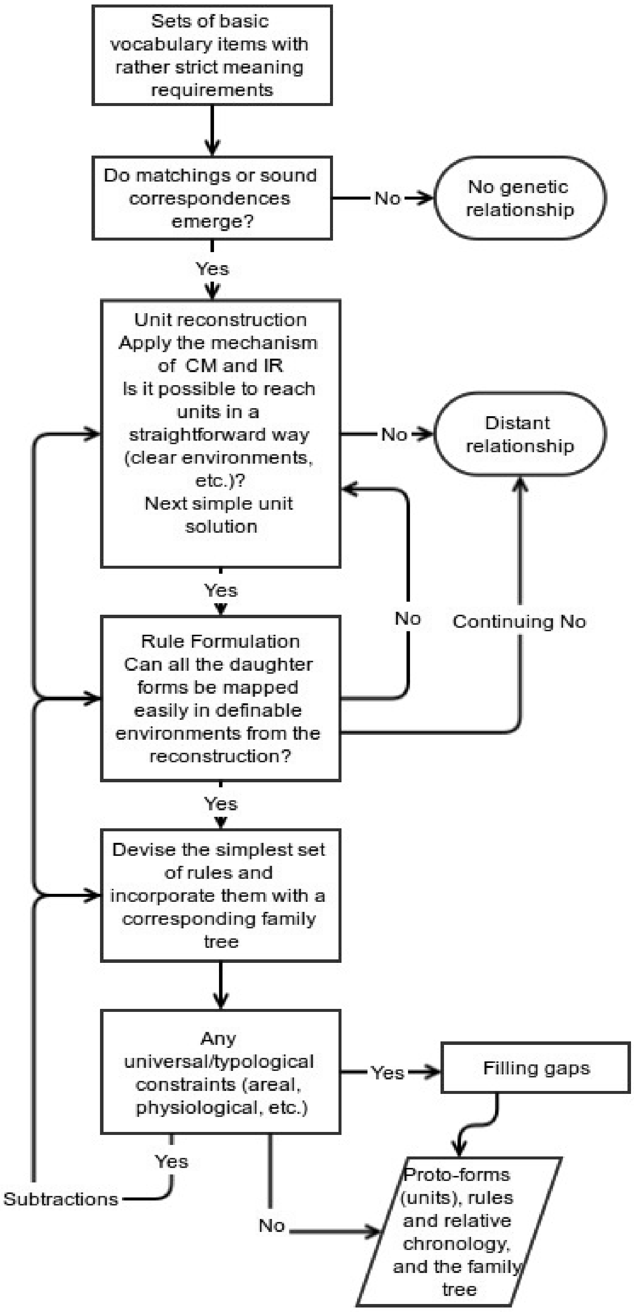

Sets of basic vocabulary items with rather strict meaning requirements
Do matchings or sound correspondences emerge?
No
No genetic relationship
Yes
Unit reconstruction Apply the mechanism of CM and IR Is it possible to reach units in a straightforward way (clear environments, etc.)? Next simple unit solution
No
Distant relationship
Yes
No
Continuing No
Rule Formulation Can all the daughter forms be mapped easily in definable environments from the reconstruction?
Yes
Devise the simplest set of rules and incorporate them with a corresponding family tree
Any universal/typological constraints (areal, physiological, etc.)
Yes
Filling gaps
Yes
Subtractions
No
Proto-forms (units), rules and relative chronology, and the family tree

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

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 in the least number of most natural changes, also referred to as *Occam's Razor*.

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 the absolute proof of relatedness and goes on to stress that irregular grammatical forms are the best evidence for establishing a *common 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. A 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 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 of travel over water or land in Latin and Armenian. So, the original meaning of **pont* is reconstructed as a general word for travel. In English, *little* and *small* are different (roughly synonymous) lexical items, whereas in Swedish the cognate forms *liten* and *små* are inflectional forms of the same lexical item (*liten*, *litet*, *lilla*, *lille* are singular forms and *små* is plural).[10] 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 against 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

[10] This example is given by Lars Borin.

an example of typological check against the reconstructed consonant system. The PIE consonant inventory which was once the most widely accepted had 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 reject Grimm's law 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 been discredited 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 the *f* in *fraternal* is not a native word but was 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 a intermediary common ancestor which has undergone this particular linguistic change and all the daughter languages of the 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 unexplainable reflexes which 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 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 such as apical displacement

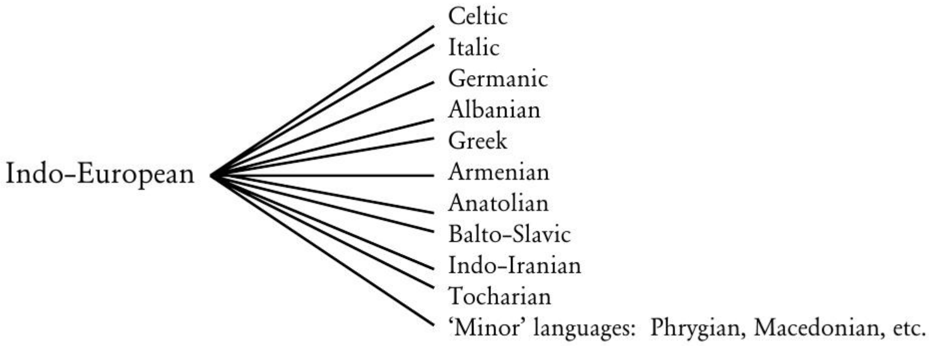


Figure 2.2: Higher-order tree of IE family from Garrett (1999).

which is still in progress (*lexical diffusion*; Chen and Wang 1975) in few languages of the South-Central Dravidian 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 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[11] and does not explicitly show information about the date of splits (Hoenigswald 1987). The date of splits can be worked out 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.

[11] A metrical tree shows branch lengths.

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. Thus, a resolved family tree need not show binary splits at all the nodes – 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 IE wave model 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 the points 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 re-conciliated with the wave model. The tree-envelope representation of Southworth 1964 is one such example which attempts at showing the subgrouping as well as the shared innovations between the subgroups. The study of lexical diffusion of *s* > *h* > *Ø* 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* > *Ø* 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.

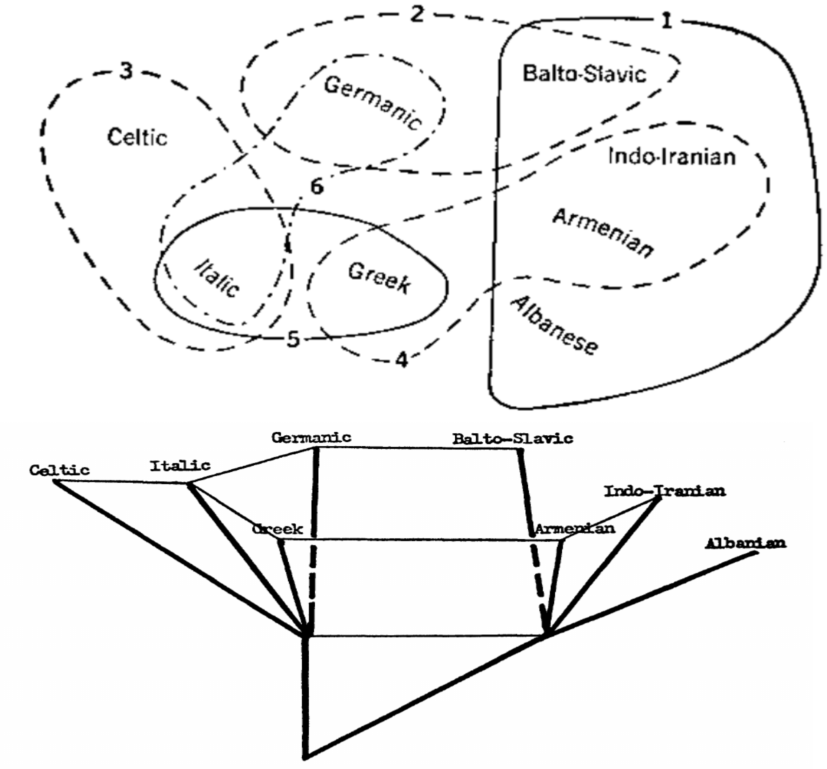
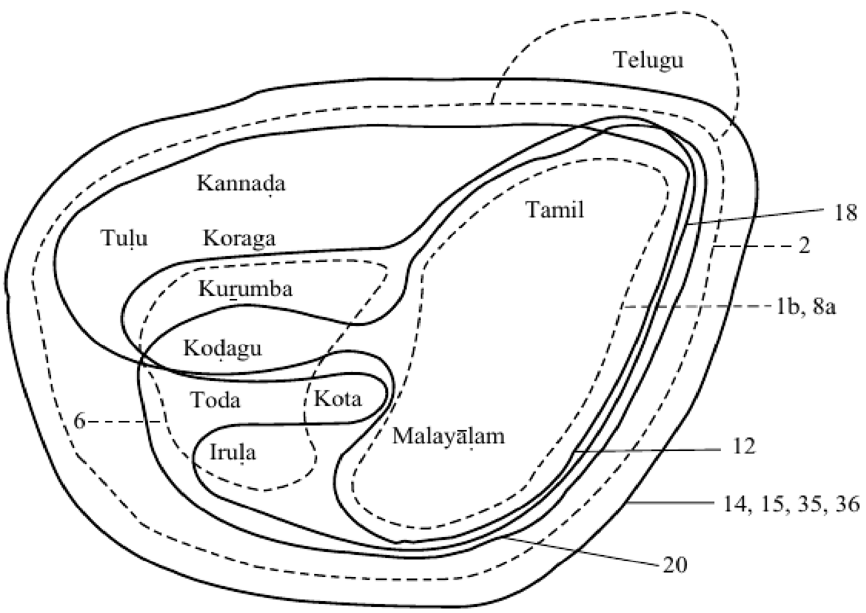


Figure 2.3: Indo-European isoglosses (Bloomfield 1935: 316) and the corresponding tree-envelope representation from Southworth (1964). The numbers in isogloss figure correspond to the following features. **1.** Sibilants for velars in certain forms. **2.** Case-endings with [m] for [b^h]. **3.** Passive-voice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

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



- - - - - - 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).

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 lan-

> guages 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 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 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).

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. Just as 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

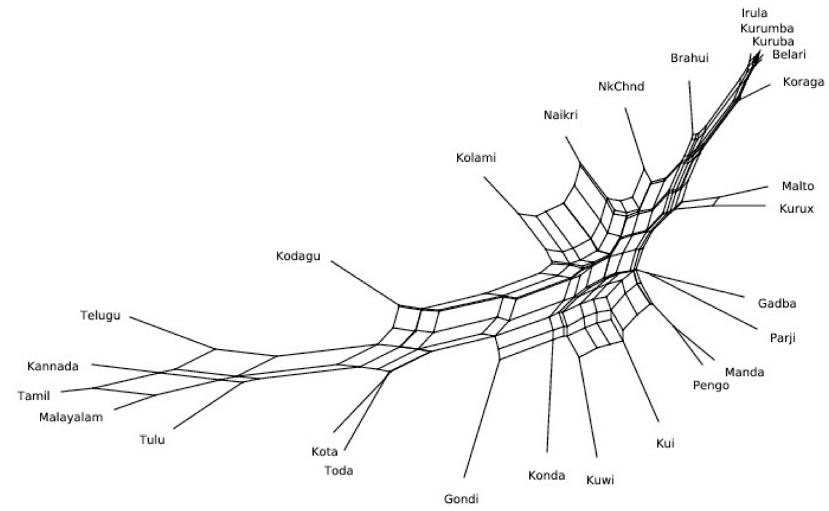


Figure 2.5: A network diagram of 28 Dravidian languages based on grammatical and phonological features (Rama and Kolachina 2013).

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 would like to 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. Unfortunately, these authors reduce the reconstruction step to a search procedure over a tree topology inferred from the comparative method. Hence, there is an inherent circularity in their method.

2.4 Alternative techniques in language classification

The standard historical linguistics textbooks list lexicostatistics and glottochronology as the alternative techniques in language classification. However none of them note that positing genetic proximity based on cognate counts and the counts of shared phonological and grammatical innovations preceded lexicostatistics. This crucial point 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. The rest of the section is on lexicostatistics and the recent classification methods that are beyond lexicostatistics. According to Wichmann (2013a), the textbooks usually portray the other methods as *discredited*.

2.4.1 Lexicostatistics

The lexicostatistical technique as introduced by Swadesh (1950) works on standardized multi-lingual word lists. In 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 100/200 -word lists. The availability of plug-and-play biological software spurred researchers to apply the 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.

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.

- *Distance-based methods*. The pair-wise cognate judgments are coded as sequence of '1's and '0's (cf. table 2.5) and the difference between the character sequences is fed to a distance based algorithm. Some popularly used distance-based algorithms are Neighbor-Joining (NJ), Unweighted Pair Group Method with Arithmetic Mean (UPGMA), Fitch-Margoliash, and FastME (Fast Minimum Evolution). All the distance methods try to optimize a criteria such as sum of the branches on the tree

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).

(tree length) or a function of the tree length. Sometimes, orthographic or phonetic-based similarity is also supplied as an input to the distance algorithms (Felsenstein 2004: chapter 11).

- *Character-based methods*. These methods also work on a sequence of characters but instead try to fit the data to a model of evolution. Maximum Parsimony is one such evolutionary principle which demands that the best tree for the data is the one which explains the change from ancestral characters to the daughter languages in least number of steps. Maximum likelihood is another method which yields a metrical tree. This method employs parameters such as branch length, frequency of change of a character from $1 \rightleftharpoons 0$, and also the differential rate of evolutions across characters as well as branches. An example of such characters is that grammatical features change at a much slower rate than lexical features; and the Anatolian branch (Hittite, Luwian, and Lycian) of the IE family are conservative (Hock and Joseph 2009: 504). The Bayesian approach includes maximum likelihood as a component and also includes a prior weight to the tree under consideration (Felsenstein 2004: chapters 1, 16, and 18).

The international consortium of scholars centered at Leipzig[12] applied Levenshtein distance for triangulating the *urheimat* (homeland) of language families, dating of the world's languages, and language classification. The Auckland group[13] has applied Bayesian techniques to various issues such as dating of PIE and Proto-Austronesian, the populating chronology of Pacific islands, and rates of evolution of typological universals.

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

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

[13] http://language.psy.auckland.ac.nz/austronesian/

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 vigorously (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).

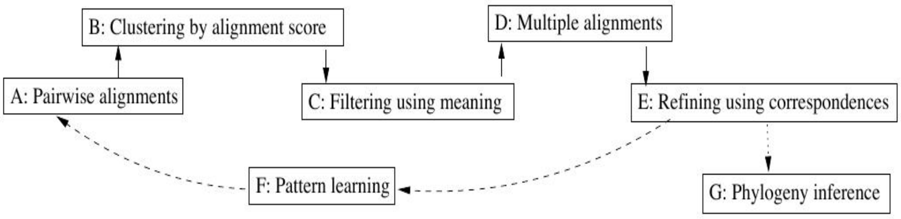

B: Clustering by alignment score
D: Multiple alignments
A: Pairwise alignments
C: Filtering using meaning
E: Refining using correspondences
F: Pattern learning
G: Phylogeny inference

Figure 2.6: A pipeline for a language classification system.

One can easily see that pair-wise 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.

2.6 Tree evaluation

In this section, various tree comparison and a reticulation measures are described. The aim of this section is to provide a summary of various tree comparison measures which are used for evaluating language classification systems. 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. This section also provides a description of a reticulation measure called δ. 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 of the amount of deviation of 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 next 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:
$$d_{ij} + d_{kl} = d_{ik} + d_{jl} \geq d_{il} + d_{jk} \tag{2}$$

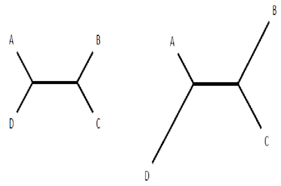


Figure 2.7: Additive trees for a quartet

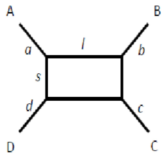


Figure 2.8: Reticulate quartet

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 right 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 consists of the comparative method and its demonstration to a medium-sized language family, such as Mayan or Dravidian family. For instance, Campbell (2004) has a chapter on the comparative method and illustrates the use of *shared innovation* in subgrouping of 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 chapter on reconstruction of proto-world is usually characterized as a *maverick* approach in historical linguistics. Any quantitative technique which attempts at dating the divergence time of a language into its daughter languages is bundled together with glottochronology.

Interestingly, Campbell (2004) uses the terms *glottochronology* and *lexicostatistics* interchangeably. Although both the 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 much 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 Andes to Indo-Aryan languages (McMahon and McMahon 2007) spoken in Northern India refrain from assigning dates to splits (McMahon and McMahon 2005: 177).

Given that there is such a huge criticism against the aforementioned techniques, how come there are so many posited families? Is the comparative method highly successful in positing these families? Unfortunately, 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 small island which do not belong to 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. Interestingly, many of the proposed language families in New Guinea are proposed based on cognate counts, similarities in pronouns, typological similarity or geographical similarity (Wichmann 2013b: originally from Foley 1986). The situation for South American languages is only a little better (Hammarström 2013), with many well established families, but also many relations that remain to be worked out using the comparative method (cf. Campbell 2012 for progress in this regard).

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 suspicion. The next subsection discusses the reality of linguistic reconstruction and attempts at correlating the linguistic evidence with archaeological and other kinds of evidence.

2.7.1 Non-quantitative methods in linguistic paleontology

Linguistic paleontology makes inferences on the culture, society, and ecology of prehistoric peoples based on reconstructed linguistic evidence (Hock and Joseph 2009: 481). Linguistic paleontology opens a window into the different aspects of life of prehistoric populations. Borrowed words corresponding to a technical innovation, names of places, and names of people allow historical linguists to assign a date to important linguistic changes affecting a pre-language. Migration histories also provide evidence for the split of the current languages from their ancestor. For instance, the vocabulary reconstructions of domesticated animals in PIE are taken to indicate that the PIE speakers were food-producers. The appearance of loan words and the subsequent sound changes they triggered, also allow historical linguists to assign a date to the sound change. For instance, looking into the Romani vocabulary and tracing the sources of loans provides information on the pattern of migration of Romani people into Europe.

Locating the probable geographical source of proto-language speakers is a highly debated topic. Historical evidence shows that the migrations of Germanic speakers caused the split of the Germanic ancestral language and this occurred about 2100 BP (before present). This date is considered as the antiquity of Proto-Germanic. The split date of Slavic languages is given around 1500 BP since the written records of sixth century describe the state of political affairs and geographic expansion of the Common Slavic language (Holman et al. 2011). The skepticism regarding the search for putative homelands can be summarized in the following quote of Mallory (1989: 143).

> the quest for the origins of the Indo-Europeans has all the fascination of an electric light in the open air on a summer night: it tends to attract every species of scholar or would-be savant who can take pen to hand

Sapir (1916) proposes a model for locating proto-language homelands called the *centre of gravity* model. Under this model, the homeland of a language family is the region that shows the highest amount of linguistic diversity. The homeland for a language family has the highest amount of divergence in terms of languages belonging to the oldest branches of the family since this point corresponds to the initial divergence of the language family.

The questions of dating, finding homelands, and evolution of cultural traits had been addressed from a computational perspective in recent years. A few examples of such research are:

- Holden (2002) applies maximum parsimony to show that the Bantu family's language trees reflect the spread of farming in sub-Saharan Africa.

- Jordan et al. (2009) apply Bayesian techniques to study the evolution of matrilocal residence from Proto-Austronesian. This is done by examining the evolution of matrilocal traits in the different Austronesian languages.

- Wichmann, Müller and Velupillai (2010) implement Sapir's idea, finding the area of greatest diversity based on lexical evidence and identify that area with the homeland; the approach is applied across the world's language families.

- Walker et al. (2012) apply Bayesian techniques to study the cultural evolution in the Tupian language family in Lowland South America.

- Bouckaert et al. (2012) apply Bayesian techniques to map the origins and expansions of the Indo-European language family.

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.

2.8 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 second part of the thesis describe the Automated Similarity Judgment Program (ASJP) database, World Atlas of Language Structures (WALS) database, and the Europarl parallel corpora (from European parliamentary proceedings). Thus, this chapter will focus on linguistic databases which are not listed in part II 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 state of phonological, lexical, and grammatical features (characters) across a language family. Core vocabulary databases with or without cognate judgments.
- 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 40–215 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/ before */r/, */u/, */k/, or */i/ – is coded as 2 in 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 Intercontinental Dictionary Series (IDS)

IDS is an international collaborative lexical database for non-prestigious and little known languages. The database is organized into 23 chapters 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 multigram alignments.[14]

3.1.4 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 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 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 the still unpublished results of the LWT, finding the absence of a correlation, suggesting, importantly, that borrowability is not a major contributor to lexical stability.

3.1.5 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 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.

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

3.1.6 Austronesian Basic Vocabulary Database (ABVD)

ABVD[15] (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)[16] 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 more languages enter into the database. The database was used for investigating quantitative and qualitative typological universals (Bickel and Nichols 2002).

[15] Accessed on 2nd December 2013.

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

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 ODIN

Online Database of Interlinear Text (ODIN; Lewis and Xia 2010) is an automatically extracted database from scholarly documents present on the web. The database has more than $190,000$ instances of interlinear text for more than $1,000$ languages. The database provides search facilities for searching the language data and the source of the data. The database is available for download. 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 database also allows search by syntactic trees and categories.

3.3.2 PHOIBLE

PHOnetics Information Base and LExicon (PHOIBLE)[17] 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).

3.3.3 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.

3.3.4 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[18] lists

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

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

the 47 languages, size of lexicon, and the source of data. Nothing much is known about the methodology and development of the corpus from the website of the project.

3.4 Conclusion

In this chapter, various linguistic databases are summarized. Not all of the databases have been used for language classification. 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 requirement grow on an exponential scale (*Zipf's law*). 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.

4 Summary and future work

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

4.1 Summary

Chapter 1 places the work in part II in the context of LT and gives 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.

The following papers have as their main theme the application of LT techniques to address some of the classical problems in historical linguistics. The papers Rama and Borin 2013, Rama 2013, and Rama and Borin 2014 work with standardized vocabulary lists whereas Rama and Borin 2011 works with automatically extracted translational equivalents for 55 language pairs. Most of the work 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.

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 Rama and Borin 2014 show that n-gram measures are good at internal classification whereas Levenshtein distance is good at discriminating related languages from unrelated ones. The chapter 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).

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.

Rama and Borin 2013 explores 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).

Rama and Borin 2011 can be seen as the application of LT techniques for corpus-based CHL. In contrast to the rest of papers which work with the ASJP database, in this paper, we attempt to extract cognates and also infer a phenetic tree for 11 European languages using three different string similarity measures. We try to find cognates from cross-linguistically aligned words by imposing a surface similarity cut-off.

4.2 Future work

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

- Exploiting 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. To make the most out of short word lists, skip-grams can be used as features to train linear classifiers (also string kernels; Lodhi et al. 2002) for cognate identification and language classification.

- 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 language dating paper can be extended to include the phylogenetic tree structure into the model. Currently, 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).

- Application of the recently developed techniques from CHL to digitized grammatical descriptions of languages or public resources such as Wikipedia and Wiktionary to build typological and phonological databases (Nordhoff 2012) could be a task for the future.

Part II

Papers on application of computational techniques to vocabulary lists for automatic language classification

5 ESTIMATING LANGUAGE RELATIONSHIPS FROM A PARALLEL CORPUS

Taraka Rama, 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

Abstract

Since the 1950s, Linguists have been using short lists (40–200 items) of basic vocabulary as the central component in a methodology which is claimed to make it possible to automatically calculate genetic relationships among languages. In the last few years these methods have experienced something of a revival, in that more languages are involved, different distance measures are systematically compared and evaluated, and methods from computational biology are used for calculating language family trees. In this paper, we explore how this methodology can be extended in another direction, by using larger word lists automatically extracted from a parallel corpus using word alignment software. We present preliminary results from using the Europarl parallel corpus in this way for estimating the distances between some languages in the Indo-European language family.

5.1 Introduction

Automatic identification of genetic relationships among languages has gained attention in the last few years. Estimating the distance matrix between the languages under comparison is the first step in this direction. Then a distance based clustering algorithm can be used to construct the phylogenetic tree for a family. The distance matrix can be computed in many ways. Lexical, syntactic and semantic features of the languages can be used for computing this matrix (Ringe, Warnow and Taylor 2002). Of these, lexical features are the most widely used features, most commonly in the form of *Swadesh lists*.

Swadesh lists are short lists (40–200 items) of basic senses which are supposed to be universal. Further, the words expressing these senses in a language are supposed to be resistant to borrowing. If these two assumptions hold, it follows that such lists can be used to calculate a numerical estimate of genetic distances among related languages, an endeavor referred to as *lexicostatistics*. A third assumption which was often made in the older literature was that the replacement rate of this basic vocabulary was constant and could be expressed as a constant percentage of the basic vocabulary being replaced over some unit of time (exponential decay). This third assumption has generally been abandoned as flawed and with it the body of research that it motivated, often referred to as *glottochronology*.

In lexicostatistics, the similarity between two languages is the percentage of shared cognates between the two languages in such a list. In the terminology of historical linguistics, cognates are words across languages which have descended independently in each language from the same word in a common ancestor language. Hence, loanwords are not cognates. Cognates are identified through regular sound correspondences. For example, English ∼ German *night* ∼ *Nacht* 'night' and *hound* ∼ *Hund* 'dog' are cognates. If the languages are far enough removed in time, so that sound changes have been extensive, it is often far from obvious to the non-expert which words are cognates, e.g. English ∼ Greek *hound* ∼ *kuon* 'dog' or English ∼ Armenian *two* ∼ *erku* 'two'.

In older lexicostatistical work (e.g. Dyen, Kruskal and Black 1992), cognates are manually identified as such by experts, but in recent years there has been a strong interest in developing automatic methods for cognate identification. The methods proposed so far are generally based on some form of orthographic similarity[19] and cannot distinguish between cognates on the one hand and loanwords or chance resemblances on the other. Confusingly, the word pairings or groups identified in this way are often called cognates in the computational linguistics literature, whereas the term *correlates* has been proposed in historical linguistics for the same thing (McMahon and McMahon 2005). In any case, the identification of such orthographically similar words is a central component in any automatic procedure purporting to identify cognates in the narrower sense of historical linguistics. Hence, below we will generally refer to these methods as methods for the identification of cognates, even if they actually in most cases identify correlates.

There have been numerous studies employing string similarity measures for the identification of cognates. The most commonly used measure is nor-

[19] Even though the similarity measures used in the literature all work with written representations of words, these written representations are often in fact phonetic transcriptions, so that we can say that we have a phonetic similarity measure. For this reason we will use "orthographic" and "phonetic" interchangeably below.

malized edit distance. It is defined as the minimum number of deletions, substitutions and insertions required to transform one string to another. There have also been studies on employing identification of cognates using string similarity measures for the tasks of sentence alignment (Simard, Foster and Isabelle 1993), statistical machine translation (Kondrak, Marcu and Knight 2003) and translational lexicon extraction (Koehn and Knight 2002).

The rest of this paper is structured as follows. Section 5.2 discusses related work. Section 5.3 explains the motivation for using a parallel corpus and describes the approach.

5.2 Related work

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. The measures were tested on vocabulary lists for the Algonquian language family and Dyen, Kruskal and Black (1992) Indo-European lists.

Many studies based on lexicostatistics and phylogenetic software have been conducted using Swadesh lists for different language families. Among the notable studies for Indo-European are the lexicostatistical experiments of Dyen, Kruskal and Black 1992 and the phylogeny experiments of Ringe, Warnow and Taylor 2002 and Gray and Atkinson 2003. In another study, Ellison and Kirby (2006) used intra-language lexical divergence for measuring the inter-language distances for the Indo-European language family.

Recently, a group of scholars (Wichmann et al. 2010a; Holman et al. 2008a) have collected 40-item Swadesh word lists for about two thirds of the world's languages.[20] This group uses a modified Levenshtein distance between the lexical items as the measure of the inter-language distance.

Singh and Surana (2007) use corpus based measures for estimating the distances between South Asian languages from noisy corpora of nine languages. They use a phonetics based similarity measure called *computational phonetic model of scripts* (CPMS; Singh, Surana and Gali 2007) for pruning the possible cognate pairs between languages. The mean of the similarity between the pruned cognate pairs using this measure is estimated as the distance between the languages.

[20] Their collaboration goes under the name of the *Automated Similarity Judgement Program (ASJP)* and their current dataset (in late 2010) contains word lists for 4,820 languages, where all items are rendered in a coarse phonetic transcription, even for those languages where a conventional written form exists.

Bergsma and Kondrak (2007) conduct experiments for cognate identification using alignment-based discriminative string similarity. They automatically extract cognate candidate pairs from the Europarl corpus (Koehn 2005) and from bilingual dictionaries for the language pairs English–French, English–German, English–Greek, English–Japanese, English–Russian, and English–Spanish. Bouchard-Côté et al. (2007) also use the Europarl corpus to extract cognates for the task of modeling the diachronic phonology of the Romance languages. In neither case is the goal of the authors to group the languages genetically by family, as in the work presented here. The previous work which comes closest to the work presented here is that of Koehn 2005, who trains pair-wise statistical translation systems for the 11 languages of the Europarl corpus and uses the systems' BLEU scores for clustering the languages, under the assumption that ease of translation correlates with genetic closeness.

5.3 Our approach

As noted above, automatic identification of cognates is a crucial step in computational historical linguistics. This requires an approach in which cognates have to be identified with high precision. This issue has been discussed by Brew and McKelvie (1996). They were trying to extract possible English-French translation pairs from a multi-lingual corpus for the task of computational lexicography. Two issues with the automatic methods is the presence of *false friends* and *false negatives*. False friends are word pairs which are similar to each other but are unrelated. Some examples of false friends in French and English are *luxure* 'lust' $\sim$ *luxury*; *blesser* 'to injure' $\sim$ *bless*. False negatives are word pairs which are actually cognates but were identified as unrelated. For our task, we focus on identifying cognates with a high precision – i.e., few false friends – and a low recall – i.e., many false negatives. The method requires that the word pairs are translations of each other and also have a high orthographic similarity.

Section 5.4 introduces the use of the Europarl corpus for cognate identification. We extract the cognate pairs between a pair of languages in the following manner. For every language pair, the corpus is word aligned using GIZA++ (Och and Ney 2003) and the word pairs are extracted from the alignments. Word pairs with punctuation are removed from the final set. Positive and negative training examples are generated by thresholding with a LCSR cutoff of 0.58.

The cutoff of 0.58 was proposed by Melamed (1999) for aligning bitexts for statistical machine translation. The reason for this cutoff is to prevent the LCSR's inherent bias towards shorter words. For example, the word pairs *saw/osa* and *jacinth/hyacinthe*[21] have the same LCSR of $2/3$ and $4/6$ which is counter-intuitive. If the words are identical, then the LCSR for the longer pair and the short pair are the same. A word alignment tool like GIZA++ aligns the words which are *probable translations of each other* in a particular sentence.

Given cognate lists for two languages, the distance between two languages l_a, l_b can be expressed using the following equation:

$$Dist(l_a, l_b) = 1 - \frac{\sum_i sim(l_a^i, l_b^i)}{N} \tag{3}$$

$sim(l_a^i, l_b^i)$ is the similarity between the ith cognate pair and is in the range of $[0,1]$. String similarities is only one of the many possible ways for computing the similarity between two words. N is the number of word pairs being compared. Lexicostatistics is a special case of above equation where the range of the *sim* function is $0|1$. The choice of the similarity function is a tricky one. It would be suitable to select a function which is symmetric. Another criterion that that could be imposed is $sim(x,y) \to [0,1]$ where x, y are two strings (or cognate pairs).

To the best of our knowledge, there is no previous work using these lexical similarities for estimating the distances between the languages from a parallel corpus. Section 5.4 describes the creation of the dataset used in our experiments. Section 5.5 describes the experiments and the results obtained. Finally the paper concludes with a direction for future work.

	pt	**it**	**es**	**da**	**nl**	**fi**	**fr**	**de**	**en**
sv	3295	4127	3648	12442	5568	**2624**	315 9	3087	5377
pt		10038	13998	2675	2202	**831**	6234	1245	6441
it			11246	3669	3086	**1333**	7692	1738	7647
es				3159	2753	**823**	6933	1361	7588
da					6350	**2149**	3004	3679	5069
nl						**1489**	2665	3968	4783
fi							**955**	**1043**	**1458**
fr								1545	6223
de									2206

sv : Swedish, **pt** : Portugese, **it** : Italian, **es** : Spanish, **da** : Danish
nl : Dutch, **fi** : Finnish, **fr** : French, **de** : German

Table 5.1: Number of cognate pairs for every language pair.

5.4 Dataset

The dataset for these experiments is the publicly available Europarl corpus. The Europarl corpus is a parallel corpus sentence aligned from English to ten

[21] Taken from Kondrak (2005a)

	pt	**it**	**es**	**da**	**nl**	**fi**	**fr**	**de**	**en**	**max**	**min**
sv	0.2994 0.5849 0.7321	0.2999 0.5876 0.7272	0.306 0.5869 0.7264	0.2012 0.6805 0.8127	0.2806 0.61 0.7516	0.3131 0.6215 0.7152	0.2773 0.6187 0.7496	0.2628 0.634 0.7577	0.282 0.6195 0.7424	da da da	fi pt fi
pt		0.2621 0.6147 0.7646	0.187 0.6824 0.8289	0.2944 0.5892 0.7289	0.2823 0.6102 0.7529	0.3234 0.5709 0.7109	0.2747 0.5711 0.7541	0.2783 0.5958 0.7467	0.2895 0.6008 0.7405	es es es	fi fi fi
it			0.2611 0.6137 0.7638	0.2923 0.5871 0.7321	0.2858 0.5916 0.7474	0.3418 0.5649 0.6954	0.2903 0.5725 0.7397	0.283 0.5847 0.7448	0.2802 0.6065 0.7473	es pt pt	fi fi fi
es				0.2965 0.5924 0.7298	0.2918 0.5992 0.7444	0.3265 0.5746 0.7081	0.2725 0.5799 0.7601	0.2756 0.5967 0.75	0.2841 0.6084 0.7475	it pt pt	fi fi fi
da					0.2829 0.6064 0.7518	0.3174 0.6196 0.7127	0.2596 0.6208 0.7639	0.2648 0.6164 0.7618	0.269 0.6201 0.7509	sv sv sv	fi fi fi
nl						0.3343 0.5743 0.7058	0.2452 0.6457 0.7843	0.2699 0.5971 0.765	0.268 0.6207 0.7616	fr fr fr	fi fi fi
fi							0.3369 0.5525 0.7027	0.3389 0.5817 0.7135	0.3218 0.6093 0.7072	sv sv sv	it fr it
fr								0.2734 0.5964 0.7555	0.2328 0.6505 0.7905	en en en	fi fi fi
de									0.2733 0.6082 0.749	sv sv da	fi fi fi

Table 5.2: The pairwise distances using Levenshtein Distance, Dice score, Longest Common Subsequence Ratio. The first, second and third entry in each cell correspond to Levenshtein distance, Dice and LCSR distances.

languages, Danish, Dutch, Finnish, French, German, Greek, Italian, Portugese, Spanish, and Swedish. Greek was not included in this study since it would have to be transliterated into the Latin alphabet.[22] The corpus was tokenized and the XML tags were removed using a dedicated Perl script. The next task was to create parallel corpora between all the 45 pairs of languages. English was used as the bridge language for this purpose. For each language pair, a sentence pair was included, if and only if there is a English sentence in common to each sentence. Only the first 100,000 sentence pairs for every language pair were included in these experiments.[23] Sentence pairs with a length greater than 40 words were not included in the final set.

All the languages of the Europarl corpus belong to the Indo-European language family, with one exception: Finnish is a member of the Finno-Ugric branch of the Uralic language family, which is not demonstrably related to Indo-European. The other languages in the Europarl corpus fall under three different branches of Indo-European:

1. Danish, Dutch, English, German and Swedish are Germanic languages and can be further subgrouped into North Germanic (or Scandinavian) – Danish and Swedish – and West Germanic – Dutch, English and German, with Dutch and German forming a more closely related subgroup of West Germanic;

2. French, Italian, Portuguese and Spanish are Romance languages, with the latter two forming a more closely related Ibero-Romance subgroup, joining French at the next level up in the family tree, and Italian being more distantly related to the other three;

3. Greek forms a branch of its own (but was not included in our experiment; see above).

We would consequently expect our experiments to show evidence of this grouping, including the isolated status of Finnish with respect to the other Europarl corpus languages.

5.5 Experiments

The freely available statistical machine translation system MOSES (Koehn et al. 2007) was used for aligning the words. The system also extracts the word alignments from the GIZA++ alignments and computes the conditional

[22] This is a task for the future.

[23] The experiments are computationally demanding and take few days to run.

probabilities for every aligned word pair. For every language pair, the word pairs that have an LCSR value smaller than the *cutoff* are discarded. Table 5.1 shows the number of pairwise cognates.

We experiment with three string similarity measures in this paper. Levenshtein distance and LCSR are described in the earlier sections. The other measures are *Dice* and *LCSR*. *Dice* is defined as twice the total number of shared character bigrams between two words divided by the total number of bigrams. In the next step, the normalized Levenshtein distance (NLD) between the likely cognate pairs are computed for every language pair. The Levenshtein distance between two words is normalized by the maximum of the length of the two words to account for the length bias. The distance between a language pair is the mean of all the word pairs' distances. The distance results are shown in table 5.2. *Dice* and *LCSR* are similarity measures and lie in the range of $[0,1]$.

We use these distances as input to a hierarchical clustering algorithm, UPGMA available in PHYLIP (Felsenstein 1993), a phylogeny inference package. UPGMA is a hierarchical clustering algorithm which infers a *ultrametric* tree from a distance matrix.

5.6 Results and discussion

Finnish is clearly the outlier when it comes to shared cognate pairs. This is shown in bold in table 5.1. Not surprisingly, Finnish shares the highest number of cognates with Swedish, from which it has borrowed extensively over a period of several hundred years. Table 5.2 shows the pair-wise language distances. The last column shows the language that has the maximum and minimum similarity for each language and distance.

Figures 5.1, 5.2, and 5.3 show the trees inferred on the basis of the three distance measures. Every tree has Spanish, Portugese and Italian under one subgroup, and Danish, Swedish and German are grouped together in all three trees. Finnish is the farthest group in all the trees except in tree 5.2. The closest languages are Danish and Swedish which are grouped together. Spanish and Portugese are also grouped as close relatives. The trees are not perfect: For instance, French, English and Dutch are grouped together in all the trees.

One can compare the results of these experiments with the tree inferred using Swadesh lists, e.g. the results by Dyen, Kruskal and Black (1992), which on the whole agree with the commonly accepted subgrouping of Indo-European (except that according to their results, English is equally far apart from Dutch/ German and Danish/Swedish). However, for its successful application to language subgrouping problems, Swadesh lists rely on a large amount of expert manual effort, both in the compilation of a Swadesh list for a new language[24] and in making the cognacy judgements required for the method used by Dyen, Kruskal and Black (1992) and others.

Working with corpora and automated distance measures, we are in a position both to bring more languages into the comparison, and avoiding the admitted subjectivity of Swadesh lists,[25] as well as potentially being able to draw upon both quantitatively and qualitatively richer linguistic data for the purposes of genetic classification of languages.

Instead, we compare our results with the only similar previous work that we are aware of, viz. with the tree obtained by Koehn (2005) from BLEU scores. Koehn's tree gets the two major branches of Indo-European – Germanic and Romance – correct, and places Finnish on its own. The subgroupings of the major branches are erroneous, however: Spanish is grouped with French instead of with Portugese, and English is grouped with Swedish and Danish instead of forming a group with German and Dutch.

Using corpora rather than carefully selected word lists brings noise into the comparison, but it also promises to bring a wealth of additional information that we would not have otherwise. Specifically, moving outside the putative core vocabulary, we will pick up evidence of language contact in the form of borrowing of vocabulary and historical spread of orthographical conventions. Thus, one possible explanation for the grouping of Dutch, English and French is that the first two have borrowed large parts of the vocabulary used in the Europarl corpus (administrative and legal terms) from French, and additionally in many cases have a spelling close to the original French form of the words (whereas French loanwords in e.g. Swedish have often been orthographically adapted, for example French *jus* ∼ English *juice* ∼ Swedish *sky* 'meat juice').

Some preliminary results of the experiments are given in table 2. For every language, Finnish has the least number of cognates.[26] Finnish shares the highest number of cognates with Swedish. This could be due to the large number of borrowings from one language to the other. Swedish shares the cognates in the following order: Danish, German, Dutch, English and then with Spanish, Italian, French and Portugese. Also the romance languages, French, Italian, Spanish excepting Portugese, share their highest number of cognates with each

[24] It is generally not a straightforward task to determine which item to list for a particular sense in a particular language, whether to list more than one item, etc.

[25] The Swadesh lists were originally compiled on the basis of linguistic experience and intuition about which senses should be universally available as words in languages and which words should be most resistant to replacement over time. These assumptions are only now beginning to be subjected to rigorous empirical testing by typological linguists, and it seems that both may be, if not outright false, then at least too simplistic (Goddard 2001; Evans and Levinson 2009; Haspelmath and Tadmor 2009b).

[26] The only exception is German.

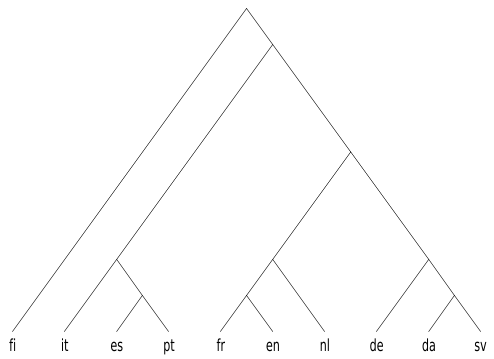


Figure 5.1: UPGMA clustering for Levenshtein distance scores

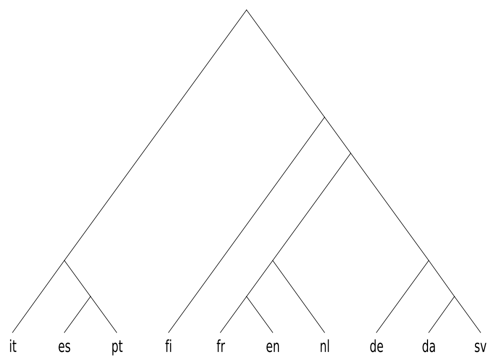


Figure 5.2: UPGMA clustering for Dice distance scores

other.

5.7 Conclusions and future work

We have presented preliminary experiments with different string similarity measures over translation equivalents automatically extracted from a parallel corpus for estimating the genetic distances among languages. The preliminary results indicate that a parallel corpus could be used for this kind of study, although because of the richer information that a parallel corpus provides, we

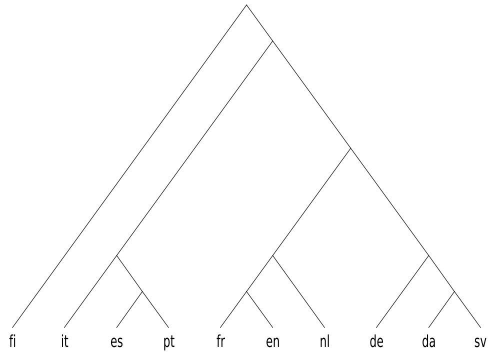


Figure 5.3: UPGMA clustering for LCSR distance scores

will need to look into, e.g., how cognates and loanwords could be distinguished. This is an exciting area for future research.

In this study, only the lexical features of the parallel corpora have been exploited, following the tradition of Swadesh list based language comparison. However, using corpora we can move well beyond the lexical level, as corpora can also be used for comparing other linguistic features. Consequently, we plan to experiment with syntactic features such as POS tags for estimating the similarity among languages. Not only the orthographic similarity but also the co-occurrence context vectors for the words could be used to estimate the similarity between translationally similar words.

6 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.

6.1 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[27] – 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.

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

[27] 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.

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.

6.2 Background and related work

6.2.1 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[28] 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.[29]

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 objectively how

[28] 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.

[29] 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.

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.

6.2.2 The ASJP database

The ASJP (Automated Similarity Judgment Program) project[30] (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 similarity between a pair

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

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).

6.2.3 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.

6.3 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 1990) 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.[31] The database contains a total of 4,169 word

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

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 6.1: 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.

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 6.1.

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 (4), 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.

$$rf^i_{ngram} = \frac{f^i_{ngram}}{\sum_{i=1}^{S} f^i_{ngram}} \tag{4}$$

Given this background, the self-entropy of the k^{th} item's *n*-gram profile can defined as in (5):

$$H^k_{item} = -\sum_{i=1}^{S} rf^i_{ngram} \cdot log(rf^i_{ngram}) \tag{5}$$

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[32]) across the families.

6.4 Results and discussion

Table 6.2 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 6.2 and the ranks given by Holman et al. (2008a)

[32] 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 100-word list sample. Both the averaging techniques correlate highly ($\rho > 0.92$, $p < 0.001$) with the weighted average.

(listed in column *H08* of table 6.2) 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.[33] 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.

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

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

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 6.2: 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.

6.5 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.[34]

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 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.

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

7 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.

7.1 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 7.2, we summarize the related work. Section 7.3 lists the contributions of this work. Section 7.4 describes the typological database, lexical database and the criteria for preparing the final dataset. Section 7.5 presents the different vector similarity measures and the evaluation procedure. The results of our experiments are given in Section 7.6. We conclude the paper and discuss the future directions in Section 7.7.

7.2 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.[35] 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 9.4). 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.

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

7.3 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?

7.4 Database

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

7.4.1 WALS

The WALS database[36] 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 7.1. 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 7.3 for the well-defined language families – Austronesian, Afro-Asiatic – given in table 7.1.

[36] 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 7.1: 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.

7.4.2 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.[37] 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 7.2. 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.

7.4.3 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 dis-

[37] 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 7.2: 10 lexical items in Russian.

crete 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).

7.5 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 7.3) followed by the evaluation measure for comparing typological distances to WALS classification. Next, we present the ASJP lexical divergence computation procedure.

Vector similarity measures

7.5.1 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

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	$\frac{\Vert\sigma_1 - \sigma_2\Vert}{2*\Vert\sigma_1\Vert+\Vert\sigma_2\Vert}$
manhattan	$\Sigma_{i=1}^{n}\lvert v_1^i - v_2^i\rvert$
chessboard	$max((v_1^i - v_2^i)\forall i \in (1,n))$
braycurtis	$\frac{\Sigma_{i=1}^{n}\lvert v_1^i - v_2^i\rvert}{\Sigma_{i=1}^{n}\lvert v_1^i + v_2^i\rvert}$
cosine	$\frac{v_1 \cdot v_2}{\Vert v_1\Vert * \Vert v_2\Vert}$
correlation	$1-\frac{\sigma_1 \cdot \sigma_2}{\Vert\sigma_1\Vert * \Vert\sigma_2\Vert}$

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	$\frac{2*\#_{\neq 0}(v_1\,\hat{}\,v_2)}{\#_{\neq 0}(v_1\&v_2)+\#_{=0}(v_1\vert v_2)+2*\#_{\neq 0}(v_1\,\hat{}\,v_2)}$
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	$\frac{\#_{\neq 0}(v_1\,\hat{}\,v_2)+\#_{=0}(v_1\vert v_2)}{\#v_1}$
yule	$\frac{2*\#_{\neq 0}(v_1-v_2)*\#_{=0}(v_1-v_2)}{\#_{\neq 0}(v_1-v_2)*\#_{=0}(v_1-v_2)+\#_{\neq 0}(v_1\&v_2)*\#_{=0}(v_1\vert v_2)}$

Table 7.3: 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, ˆ 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.

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 7.1, 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).

7.5.2 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 7.1.

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.

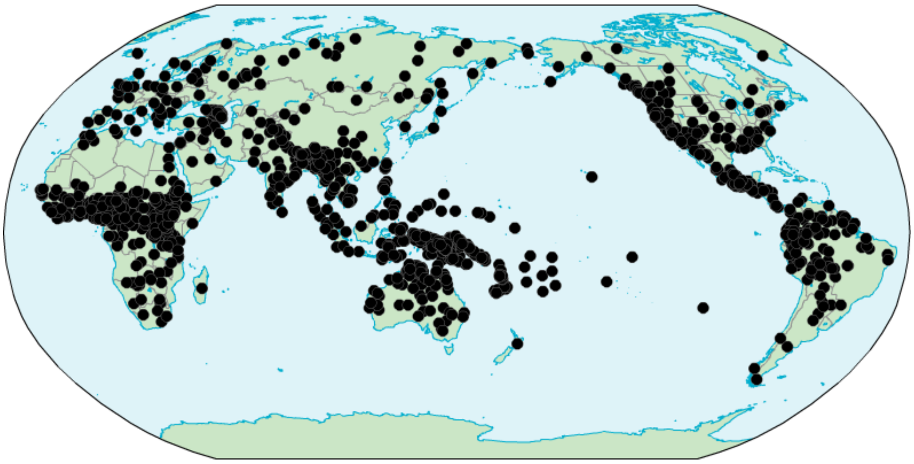


Figure 7.1: Visual representation of world’s languages in the final dataset.

7.6 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.

7.6.1 Internal classification

The point bi-serial correlation, r, introduced in section 7.5, 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 7.1. 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’.

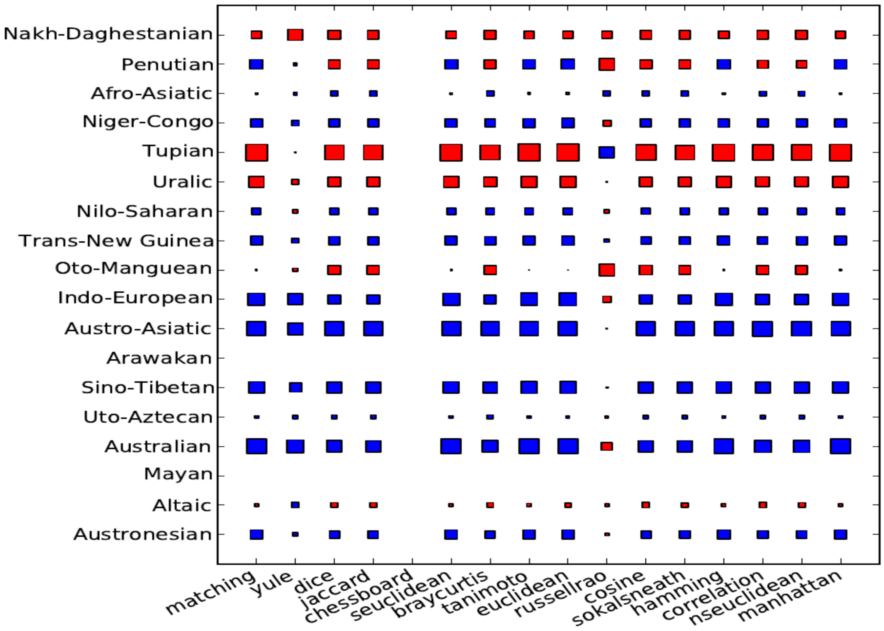


Figure 7.2: Heatmap showing the gradience of *r* across different language families and vector similarity measures.

7.6.2 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.

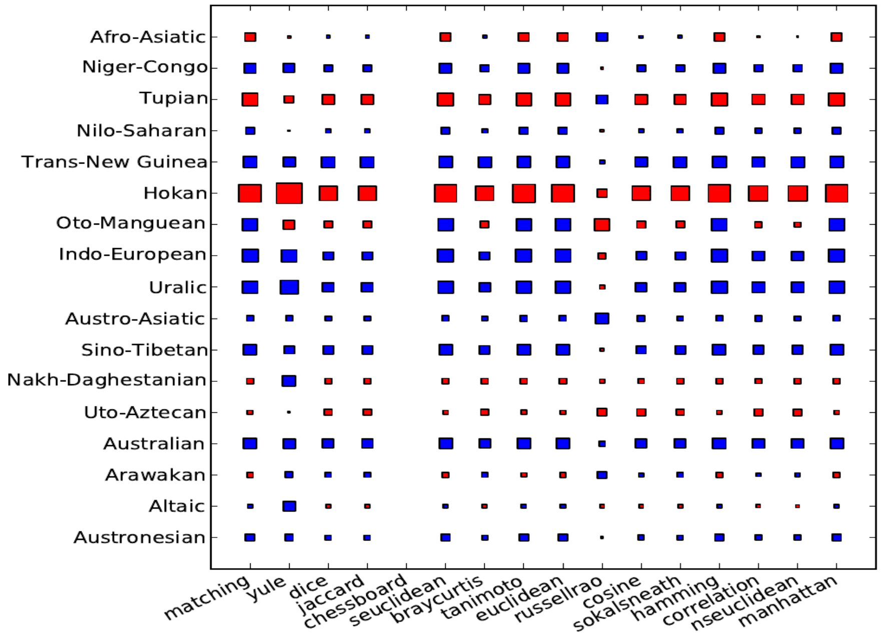


Figure 7.3: Heatmap showing the gradience of ρ across different families and vector similarity measures.

7.7 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.

8 PHONOTACTIC DIVERSITY AND TIME DEPTH OF LANGUAGE FAMILIES

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.

8.1 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.

$$t = \frac{log\ C}{2\ log\ r} \tag{6}$$

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.

8.1.1 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 and time depth as the

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

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.

8.2 Materials and Methods

8.2.1 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.

8.2.2 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 inher-

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

ited 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 (6). Eqn. 6 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):

$$t = (log\ s - log\ s_0)/2\ log\ r \tag{7}$$

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. 7 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.

8.2.3 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 A.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 8.1, 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 8.1 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 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

[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).

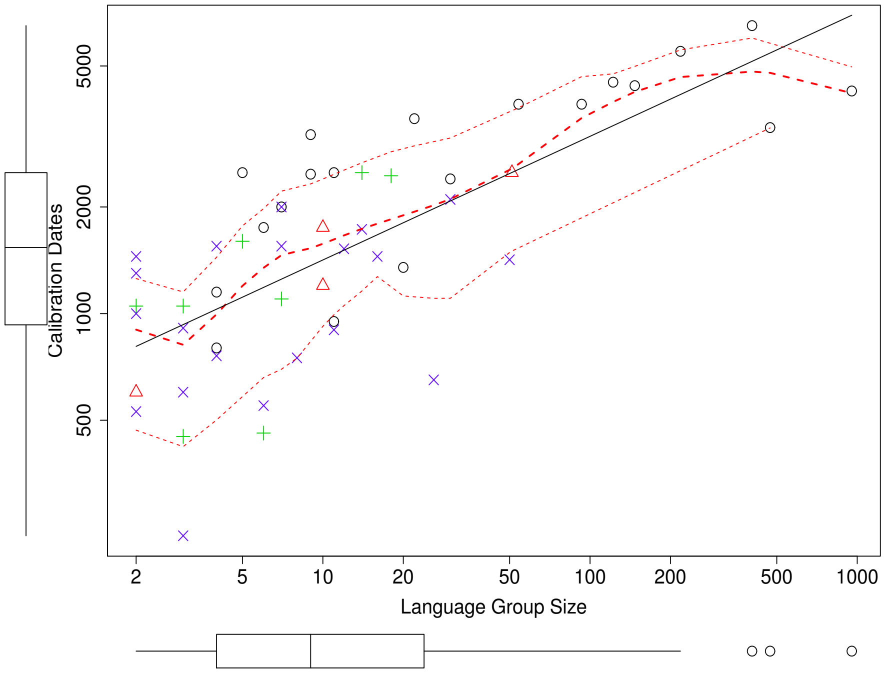


Figure 8.1: Calibration dates against the number of languages in a language group. ○s are archaeological, △s are archaeological and historical, +s are epigraphic and ×s are historical dates.

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.

8.2.4 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*.

8.2.5 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 1990) 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 ASJP-code) 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.

8.3 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 the nature of relationship between N-grams and dates. The last row in figure 8.2 shows the scatterplots of calibration dates (CD; given in table A.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

[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.

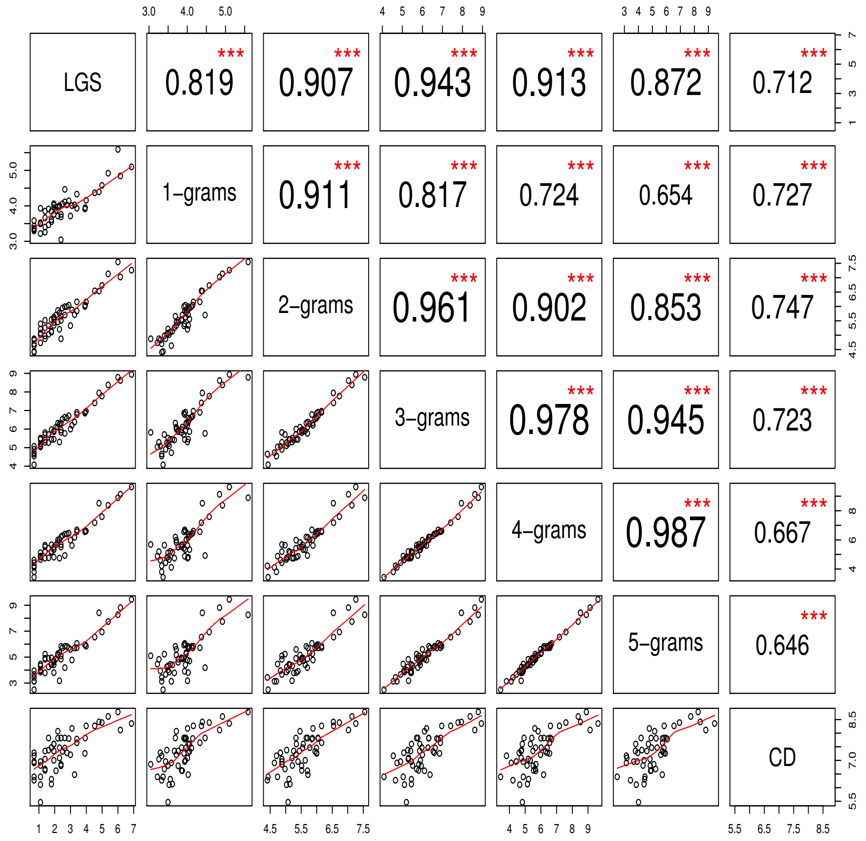


Figure 8.2: Pairwise scatterplot matrix of group size, $N-$gram diversity and date; the lower matrix panels show scatterplots and LOESS lines; the upper matrix panels show Spearman rank correlation (ρ) and level of statistical significance ($\star$). The diagonal panels display variable names. All the plots are on a log-log scale.

size, as displayed in figure 8.2. 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 8.1. 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 8.1: 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 $p < 0.001$.

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 A.1. Table 8.2 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.

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 thr-

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

ough the standard diagnostic plots, given in section A.2 – figures A.1, A.2, A.3, A.4, and A.5. The diagnostic plots of 3-gram model in figure A.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 A.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 8.3.

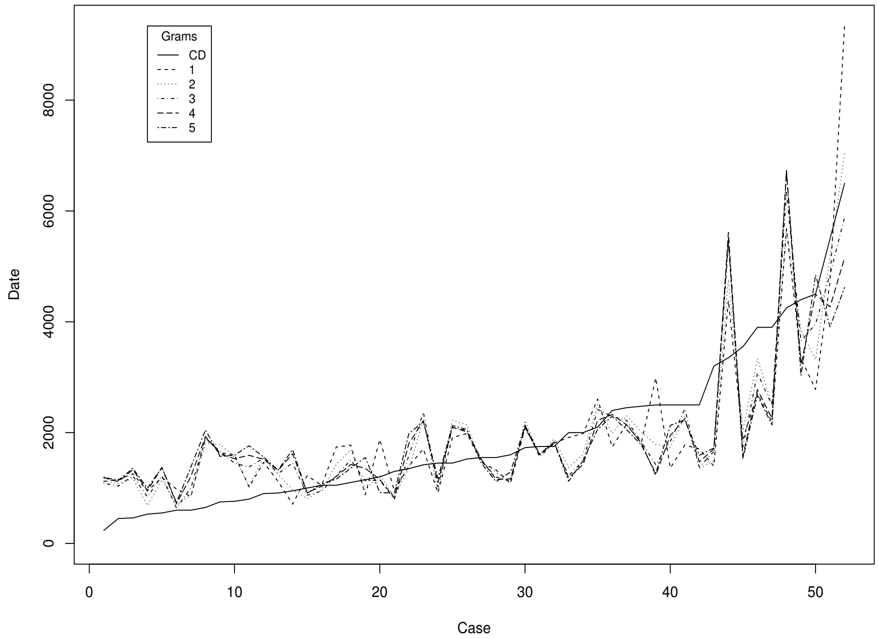


Figure 8.3: Comparing predicted dates for various n-grams

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:

$$COD = k * ASJPD + (1 - k) * NGD \qquad (8)$$

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 8.4. The horizontal axis displays the scaled k ranging from 0 to 100. Figure 8.4 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 $k = .59$.

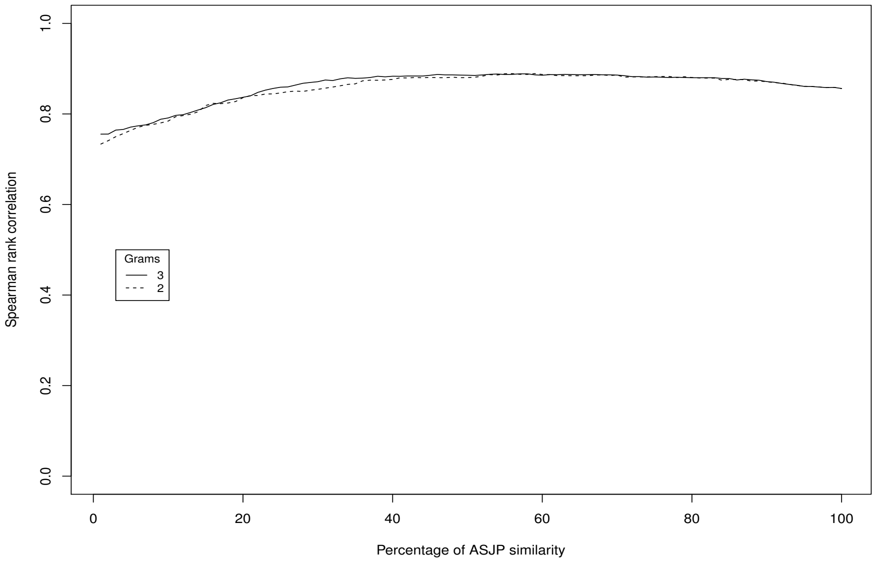


Figure 8.4: Combining ASJP with 2−grams and 3−grams: The ASJP dates are combined with 2−gram dates and 3−gram dates in different proportions ranging from 1% to 100% at an interval of 1.

The effect of mixing of 3-gram dates with ASJP dates is tested in table 8.3. Table 8.3 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.

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 8.3: A comparison of different dating methods

8.3.1 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. 8. Both the predicted and the combined dates are given in tables A.2–A.6 (section A.3). Each table presents the dates for all language families belonging to a geographical area – as defined in section 8.2. 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. 8. 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 A.2–A.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 A.3). Both models predict similar dates for Eurasian language families. The dates for languages of Pacific area is given in table A.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 A.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 A.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.

8.4 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 A.3.

9 EVALUATION OF SIMILARITY MEASURES FOR AUTOMATIC LANGUAGE CLASSIFICATION.

Rama, Taraka and Lars Borin 2014. Comparison of string similarity measures for automated language classification. Under review.

9.1 Introduction

Historical linguistics, the oldest branch of 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.[43] Although, historical linguistics has a *curious parallel origins* with biology (Atkinson and Gray 2005), unlike the biologists, main-stream historical linguists have never been enthusiastic about using quantitative methods for the discovery of language relationships or predicting the structure of a language family. However, an earlier shorter period of enthusiastic application of quantitative methods marked by Swadesh (1950) ended with the critique of Bergsland and Vogt 1962. The field of computational historical linguistics did not receive much attention until the beginning of 90's with the exception of two note-worthy doctoral dissertations: Embleton 1986; Sankoff 1969.

In traditional lexicostatistics, as introduced by Swadesh (1952), distances between languages are based on human expert cognacy judgments of items in standardized word lists, e.g., the Swadesh lists (Swadesh 1955).[44] 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[45]

[43] The Indo-European family is a classical case of the successful application of comparative method which establishes a tree relationship between the most populous languages of the world

[44] Gilij (2001) is one of the earliest known attempts at using core vocabulary for positing inter-language relationships in Americas.

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

represents such an approach, where automatically estimated distances between languages are input to 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 before hand. 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 pair-wise 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 UPGMA[46] to infer a tree structure for the set of languages. One such attempt by Swadesh (1950), even before the discovery of the first clustering algorithm: UPGMA, for Salish languages is reproduced in figure 9.1.

In the terminology of historical linguistics, cognates are related words across languages and can be traced back to the proto-language. Cognates are identified through regular sound correspondences. Usually 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 of $k^{w}ek^{w}lo-$. Usually, the cognate judgments are obtained from the expert historical linguist's judgments.

The crucial element in these approaches is the method used for determining the overall similarity between two word lists.[47] 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 transform one string into the other. A modified LD is used by the ASJP consortium, as reported in their publications (e.g., Bakker et al. 2009, Holman et al. 2008b, and Holman et al. 2008a). We describe the related work in the next section.

[46] `http://en.wikipedia.org/wiki/UPGMA`

[47] At this point, we use word list and language interchangeably. Strictly speaking, a language, identified by its ISO 639-3 code, can have as many word lists as the number of dialects.

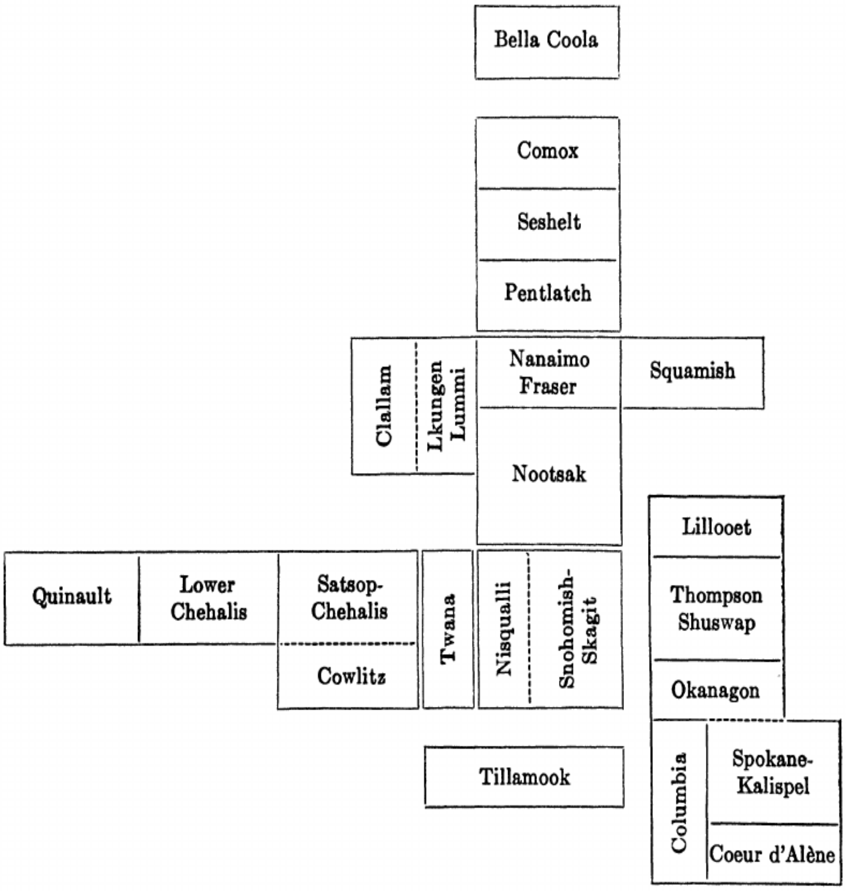


Figure 9.1: Salish box-diagram from Swadesh 1950.

9.2 Related Work

In this section, we survey the earlier work in cognate identification and distributional similarity measures for computing inter-language distances. The tasks of cognate identification and tree-inference are closely related tasks in historical linguistics. Taking each task as 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 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.[48] The inter-language distance approach to tree-building, is incidentally straight-forward and comparably accurate in comparison to the computationally intensive Bayesian-based tree-inference approach of Greenhill and Gray 2009.

The inter-language distances are either an aggregate score of the pair-wise 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.

9.2.1 Cognate identification

The task of identifying the genetically related words – cognates – is very well explored in computational linguistics. 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.

He developed a string matching algorithm based on articulatory features (called ALINE) for computing the similarity of a word pair. Although the algorithm is linguistically sound, it requires a International Phonetic transcription (IPA) transcribed input. ALINE was further evaluated against machine learning algorithms such as Dynamic Bayesian Networks and Pair-wise HMMs for automatic cognate identification (Kondrak and Sherif 2006). Even though, the approach is technically sound, it suffers due to the bare-boned phonetic transcription used in Dyen, Kruskal and Black's Indo-European dataset.[49]

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 too language specific to be generalized for the rest of Indo-European family.

In another work, Ellison and Kirby (2006) use Scaled Edit Distance (SED)[50] for computing intra-lexical similarity for estimating language distances based on the dataset of Indo-European languages prepared by Dyen, Kruskal and

[48] This idea is quite similar to the famous paradigm of Expectation-Maximization in machine learning field. Kondrak (2002b) uses this paradigm for extracting sound correspondences from pair-wise word lists for the task of cognate identification. The recent paper of Bouchard-Côté et al. (2013) employs a feed-back procedure for the reconstruction of Proto-Austronesian with a great success.

[49] Available on `http://www.wordgumbo.com/ie/cmp/index.htm`.

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

Black (1992). The distance matrix is then given as input to the NJ algorithm as implemented in PHYLIP package (Felsenstein 2002) to infer a tree for 87 languages of Indo-European family.

Petroni and Serva (2010) use a modified version of Levenshtein distance for inferring the trees of 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 normalized LD (LDN) can then be used in computing distances between pairs of languages.

There are at least two ways of computing inter-language distances: LDN and LDND. Only pairs of words occupying the same slot in two lists are compared, i.e., words expressing the same sense in the two languages. LDN for two languages is computed as the mean of the normalized LD of all compared word pairs. To compensate for chance similarity, LDN is further normalized by the mean of all $N(N-1)/2$ words to yield LDND, employed by Holman et al. (2008a) for automatic language classification.

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 the languages belonging to a same family from the languages of other families. LDND was designed to make sure that the chance resemblance, similarity in the phonemic inventory between unrelated languages do not influence the distance between them.

9.2.2 Distributional measures

Huffman (1998) compute pair-wise 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 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.

Taking cue from 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). These authors compute NJ and Minimum

Evolution trees[51] for LDN as well as LDND distance matrices. They compare the inferred trees to the classification given in Ethnologue (Lewis 2009) using two different tree similarity measures: Generalized Robinson-Fould's distance (GRF; A generalized version of Robinson-Fould's [RF] distance) and Generalized Quartet distance (GQD). GRF and GQD are specifically designed to account for the polytomous nature – a node having more than two children – of the Ethnologue trees. Finally, Huff and Lonsdale (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 ALINE over LDND.

However, 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, so far there has to our knowledge been no study to compare different string similarity measures on the same dataset in order to determine their relative suitability for linguistic classification.[52] In this paper we compare various string similarity measures[53] for the task of automatic language classification. We evaluate their effectiveness through distinctiveness measure and comparing them to the language classifications provided in WALS (World Atlas of Language Structures; Haspelmath et al. 2011) and Ethnologue.

9.3 Contributions

In this article, we ask the following questions:

- Out of the numerous string similarity measures given in section 9.5:
  - Which is the best suited for the tasks of distinguishing related lanugages from unrelated languages?
  - Which is the best suited for the task of internal language classification?
  - Is there a statistical procedure for determining the best string similarity measure?
- What is the best way to length normalize a string similarity measure?[54]

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

[52] 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.

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

[54] Marzal and Vidal (1993) propose an alternate normalization based on the length of the

9.4 Database and expert classifications

In this section, we describe the ASJP database and the two expert classifications: WALS and *Ethnologue*.

9.4.1 Database

The ASJP database offers an attractive alternative to corpora as the basis for massive cross-linguistic investigations. The ASJP effort began with a small dataset of 100-word lists for 176 languages. These languages belong to more than 23 language families, as defined in WALS (Haspelmath et al. 2011). Since Brown et al. (2008), the ASJP database has been going through an expansion, to include in its latest version (v. 14) more than 5500 word lists representing well over one half of the languages of the world Wichmann et al. 2011b). Because of the findings reported by Holman et al. (2008a), most of the added word lists have aimed to cover only the 40-item most stable Swadesh subset, and not the full 100-item list.

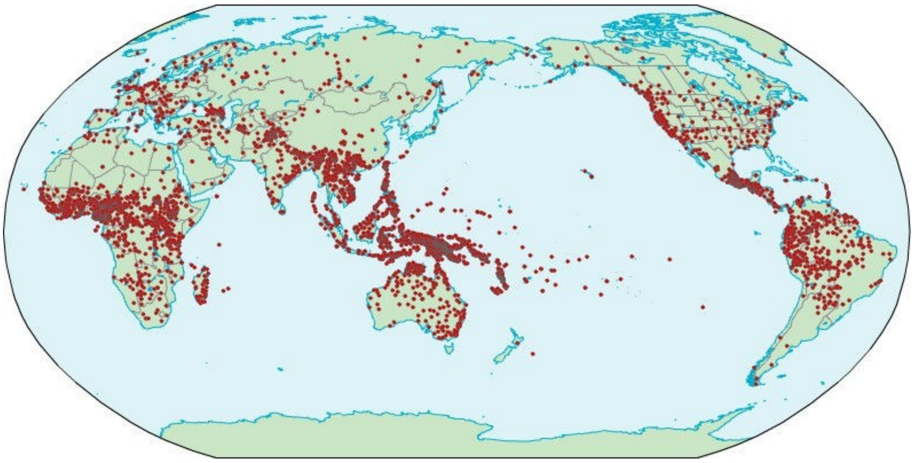


Figure 9.2: Distribution of languages in ASJP database (version 14).

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 three 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

editing path. Kondrak (2005b) tests this claim on three different datasets and finds that there is no significant difference between the two normalizations.

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 version 12 (Wichmann et al. 2010c).[55] This database comprises of reduced (40–item) Swadesh lists for 4169 word lists. 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. This leaves a dataset with 3730 word lists. It turned out that 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 excluded from the analysis.

All the experiments reported in this paper were performed on a subset of version 14 of the ASJP database whose distribution is shown in figure 9.2.[56] The database has 5500 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. Also, any word list containing less than 28 words (70%) 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. The final dataset for our experiments has 4743 word lists for 50 language families.

WALS. WALS classification is a two-level classification where each language belongs to a genus as well as 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, Indic languages are classified as a separate genus from Iranian languages although, it is quite well known that both Indic and Iranian languages share a common proto-Indo-Iranian ancestor.

Ethnologue. Ethnologue classification is a multi-level tree classification for a language family. Ethnologue classification was originally given by missionaries and is very opportunistic in the inclusion of languages or genetic relatedness. The highest node in a family tree is the family itself and languages form the lowest node. A internal node in the tree is not necessarily binary and can

[55] The original python program was kindly provided by Søren Wichmann. We modified the program to handle the ASJP modifiers.

[56] Available on 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 9.1: Distribution of language families in ASJP database. WN and WLs stands for WALS Name and Word Lists.

have more than two branches emanating from it. For instance, the Dravidian family has four branches emanating from the top node.

9.5 Methodology

In this section, we describe the various string and distributional similarity measures which form the basic component in calculating the pair-wise language distances. As mentioned earlier, string similarity measures work at the level of individual words whereas distributional measures work at the level of word-lists.

9.5.1 String similarity

We describe the different string similarity measures used in this paper.

- *IDENT* returns 1 if the two words are equivalent else returns 0.

- *PREFIX* returns the length of the common prefix divided by the length of the longer word.
- *DICE* is defined as the number of shared segment bigrams divided by the total number of bigrams in both the words.
- *LCSR* is defined as the length of the longest common subsequence divided by the length of the longer word.
- *TRIGRAM* is defined as twice the number of shared trigrams divided by the sum of the shared trigrams.
- *XDICE* is defined similar to DICE where a bigram is a trigram with the middle letter removed.
- Jaccard's index, *JCD*, is a set cardinality measure. It is defined as the ratio of the cardinality of the intersection of the bigrams between the two words to the cardinality of the union of the sets of bigrams between two words.

Each M_{sim} is converted to its distance counterpart by subtracting the score from 1.0. Lin (1998) investigates three distance to similarity conversion techniques and motivates the results from an information-theoretical point. In this article, we do not investigate this rather, we stick to the traditional conversion technique. Note that this conversion can sometimes result in a negative distance which is due to the double normalization involved in LDND. A suffix "D" is added to each measure to indicate the LDND variant.

9.5.2 N-gram similarity

This section describes the distributional similarity measures originally developed for automatic language identification in a multi-lingual document. This line of work started with the heavily cited paper of Cavnar and Trenkle 1994 who use character *n*-grams for text categorization. They observe that the frequency of character *n*-grams for a particular document category follows a Zipfian distribution. The rank of character *n*-grams vary for different categories and documents belonging to the same category have similar Zipfian distributions. Based on this idea, Dunning (1994) postulates that the character *n*-grams for a language has its own signature distribution. Comparing the signatures of two languages can yield a single point distance between the languages under comparison. The comparison procedure is usually accomplished through the use of a distributional similarity measure (Singh 2006). The following steps are followed for extracting the phoneme *n*-gram profile for a language.

For a language, the *n*-gram extraction procedure is as follows:

- A *n*-gram is defined as the consecutive phonemes in a window of *N*. The value of *N* usually ranges from 3 to 5.
- All the unigrams, bigrams and trigrams 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.
- Only the top *G* *n*-grams are retained and the rest of them are pruned. Usually *G* is fixed at 50.

For every pair of language, the *n*-gram profiles are compared using Out-of-Rank measure, Jaccard's index, Dice distance, Overlap distance, Manhattan distance and Euclidean distance. The distances are explained below:

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 language.
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.

One possible set of measures, based on *n*-grams, and could be included are information theoretic based such as cross entropy and KL-divergence. These measures have been well-studied in natural language processing systems such as machine translation, natural language parsing, sentiment identification, and also in automatic language identification. The probability distributions required for using these measures are usually estimated through maximum likelihood estimation which require a fairly large amount of data. While replicating the original ASJP experiments on the version 12 ASJP database, we tested if the above distributional measures, [1–4] perform as well as LDN. Unfortunately, the results are on the discouraging side of the spectrum and we do not repeat the experiments on the version 14 of the database. One main reason for this result is the relatively small size of ASJP word list. The relatively small word list size provides a poor estimates of the true language signatures. The next section describes the three different evaluation measures for comparing the different string similarity measures.

9.6 Evaluation measures

In this section, we describe the three different evaluation measures for ranking the string similarity measures given in section 9.5.

1. The first measure *dist*, originally given by Wichmann et al. (2010a) to test if LDND performs better than LDN in the task of distinguishing related languages from unrelated languages.
2. The second measure, *RW*, is a special case of pearson's **r** – called point biserial correlation (Tate 1954) – for computing the agreement between WALS's two-level classification and the intra-family pair-wise distances.
3. The third measure, γ, is related to the Goodman and Kruskal's Gamma (1954) which measures the strength of association between two ordinal variables. In this paper, it is used to compute the level of agreement between the Ethnologue classification and the pair-wise intra-language distances.

9.6.1 Dist

Initially, we define a function fam which takes a language name and returns the family to which the language belongs.

When,$fam(i) = fam(j)$

$$d_{in} = \frac{1}{\sum_{i=j} 1} \sum_{i \neq j} d_{i,j} \tag{9}$$

When,$fam(i) \neq fam(j)$

$$d_{out} = \frac{1}{\sum_{i \neq j} 1} \sum_{i \neq j} d_{i,j} \tag{10}$$

$$sd_{out} = stdev(d_{i,j}) \tag{11}$$

$$dist = \frac{d_{in} - d_{out}}{sd_{out}} \tag{12}$$

The *dist* measure for a family consists of two components: the pair-wise distances inside a family (d_{in}) and the pair-wise distances from each language in the family to the rest of the language families. A comparatively higher dist value suggests that a measure is particularly resistant to random similarities between unrelated languages and performs well at distinguishing elements from a genealogical unit from the rest of the dataset. The resistance of a string similarity measure to other language families is reflected by the value of sd_{out} which denotes the standard deviation of d_{out}.

9.6.2 Correlation with WALS

The WALS database has classification at three levels. The top level is the family of the language, second level is the genus and the lowest level is the language itself. Two languages from different genera and same family have a distance of 2 and a distance of 1 if they are in the same genus. This allows us to define a inter-language distance matrix between the languages in a family. The WALS distance matrix can be compared to the distance matrices of any measure using Pearson's **r**. The families with a single genus have no correlation and the corresponding row is left blank.

9.6.3 Agreement with Ethnologue

Given a pair-wise distance-matrix d of size $N \times N$, 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. Find 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, j, 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 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.

3. For a partition $\{i, j \mid k\}$, the distance matrix d is said to agree completely when the following conditions are satisfied:

$$d_{ij} < d_{ik} \tag{13}$$

$$d_{ij} < d_{jk} \tag{14}$$

   A triplet that satisfies these conditions is counted as a concordant comparison, C; else it is counted as disconcordant comparison, D.

4. Steps 2 and 3 are repeated for all the $\binom{N}{3}$ triplets to yield γ for a family defined as $\gamma = \frac{C-D}{C+D}$.

At this point, one might wonder about the decision for not using a off-the-shelf tree-building algorithm to infer a tree and compare the resulting tree with the Ethnologue classification. Although both the works of Pompei, Loreto and Tria 2011; Huff and Lonsdale 2011 compare their trees to Ethnologue using cleverly crafted tree-distance measures, they do not compare their distance matrices directly. A direct comparison of a family distance matrix to the family's Ethnologue tree removes the errors induced by the tree inference algorithm. The results of Wichmann et al. 2011a, suggest that the GE measure shows a high negative correlation with the GRF and GQD for the LDND distance matrix computed from ASJP database (v. 14) families.

9.7 Item-item vs. length visualizations

A distance measure such as LDND has two components:

- The average LDN computed between pair-wise lexical items for the same meaning.
- The average LDN computed between pair-wise lexical items for different meaning-meaning pair.

Although, there have been multitude of publications involving LDND, no effort has been put to look the base components of LDND through a magnifying glass. It is quite interesting to see the overall distribution between LD and the lengths of the lexical items under comparison. We proceed to see the variation of the pair-wise LDs vs. the pair-wise word lengths for lexical items sharing the same meaning and different meanings. We select a small (31 word lists) but well-studied language family called Dravidian family, spoken in South Asia (comprising of modern day India, Pakistan, and Nepal), for this study.

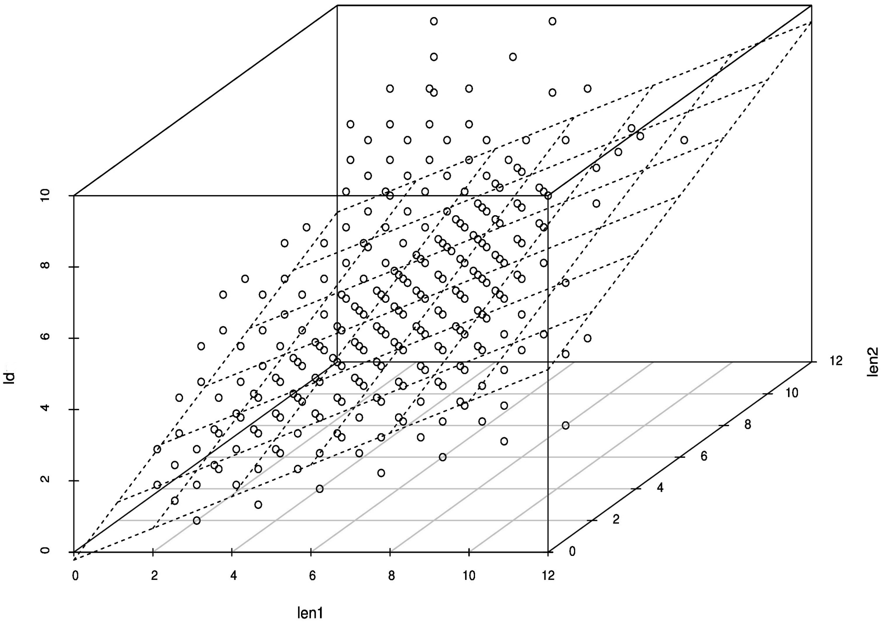


Figure 9.3: 3D scatterplot for Dravidian language family and same items. There are more than 21,000 points in this plot.

We compute the LD between all the word pairs for a same meaning. We plot the corresponding LD and the word-length pairs on a three-dimensional plot as shown in figure 9.3. We repeat the exercise now for the word-pairs for different meaning pairs and plot them in figure 9.4. A 3-dimensional scatterplot would has the advantage of showing the distribution of pair-wise LD distances against the length of the word-pairs. In the next step, we fit a linear regression plane to the distribution. This plane would show the number of points which fall below or above the regression plane. The regression shown in figure B.1 is highly significant and has an adjusted-R^2 value of 0.2554 ($p < 2.2e - 16$). The multiple regression for different meaning-meaning pairs shows an adjusted-R^2 value of 0.4953 ($p < 2.2e - 16$). These multiple regressions support the hypothesis that there exists a linear relationship between pair-wise lengths and LD.

Given that, the multiple regressions support a linear function of pair-wise lengths, it would be interesting to plot a scatterplot of LD vs. average length and LD vs. maximum length for the above two datasets. We employ the hexagonal binning technique for showing the huge number of points. The size of a bin is depicted through the color intensity and a color legend shows the number of points in a bin. The results of this technique is shown in the figures B.1,

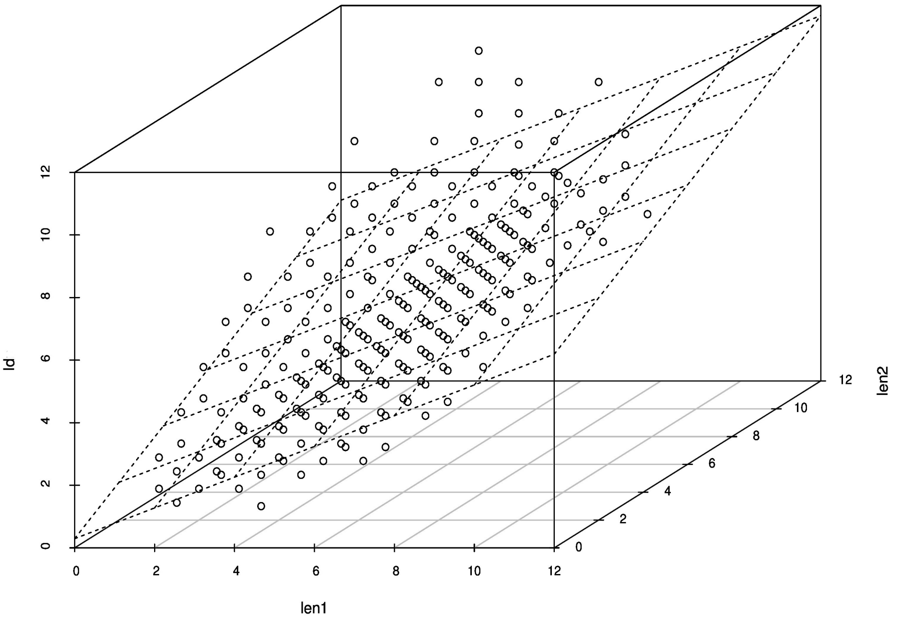


Figure 9.4: 3D scatterplot for Dravidian language family and different items. There are about 1 million data points resulting from this exercise. So, we randomly show 22,000 points in this plot.

B.2, B.3, and B.4. Each of the hexagonal plot is plotted using 100 bins. Average length vs. LD is more dispersed in same as well as different meaning pairs. The maximum length normalization seems to follow LD closely in both the plots. We take this as visualization as a support for preferring maximum length normalization over average length normalization.

The results of comparison of the various string similarity measures is described in the next section.

9.8 Results and discussion

In this section we give the results of our experiments in table 9.2. We only report the average results for all measures across the families listed in table 9.1. We further check the correlation between the three measures by computing a Spearman's ρ. The pair-wise ρ is given in table 9.3. The high correlation value of 0.95 between RW and γ suggests that all the measures agree roughly on the task of internal classification.

The average scores in each column suggests that the string similarity mea-

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 9.2: 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 9.3: Spearman's ρ between γ, RW and Dist

sures show a variable 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.[57] 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

[57] http://en.wikipedia.org/wiki/Sign_test

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 γ, Dist and RW for each family in tables B.1, B.2, and B.3. In each of these tables, only those measures which are above the average scores, from table 9.2, are reported. A empty row in table B.1 implies that all triplets in Ethnologue tree are ties and; in table B.3, implies that there is a single genus in that family.

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 9.4. Based on the computational complexity and the significance scores, we propose that JCDD be used for step 1 and measure like LDN be used for internal classification.

9.9 Conclusion

We conclude the article by pointing that this is the first known attempt at applying more than 20 similarity measures for more than half of the world's lan-

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 9.4: Computation complexity for 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 $C = l(l-1)/2$

guages. We examine various measures at two levels, namely, distinguishing related from unrelated languages and internal classification of related languages.

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.

Anttila, Raimo 1989. *Historical and comparative linguistics*. Volume 6 of *Current Issues in Linguistic Theory*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

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.

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.

Beekes, Robert Stephen Paul 1995. *Comparative Indo-European linguistics:*

An introduction. Amsterdam and Philadelphia: John Benjamins Publishing Company.

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.

Bergsma, Shane and Grzegorz Kondrak 2007. Alignment-based discriminative string similarity. *Proceedings of the 45th annual meeting of the association of computational linguistics*, 656–663. Prague, Czech Republic: Association for Computational Linguistics.

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.

Bloomfield, Leonard 1935. *Language*. London: Allen, George and Unwin.

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 2013a. For better or for worse? Going beyond short word lists in computational studies of language diversity. Presented at Language Diversity Congress, Groningen.

Borin, Lars 2013b. 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.

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.

Bouchard-Côté, Alexandre, Percy Liang, Thomas L. Griffiths and Dan Klein 2007. A probabilistic approach to diachronic phonology. *Empirical methods in natural language processing*.

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.

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.

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.

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, US.

Chen, Matthew Y. and William S-Y. Wang 1975. Sound change: actuation and implementation. *Language* 51: 255–281.

Collinge, Neville Edgar 1985. *The laws of Indo-European*. Volume 35. 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*. 4. 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*. 2. New York: Oxford university press.

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, Rebecca Hetherington, James McElvenny and Virginia Dawson 2013. World phonotactics database. Department of Linguistics, The Australian National University. http://phonotactics.anu.edu.au.

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.

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.

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.

Evans, Nicholas and Stephen C. Levinson 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32: 429–492.

Evans, Steven N., Don Ringe and Tandy Warnow 2006. Inference of divergence times as a statistical inverse problem. *Phylogenetic methods and the prehistory of languages. McDonald institute monographs*, 119–130.

Fellbaum, Christiane 1998. *WordNet: An electronic database*. Cambridge, Massachusetts: MIT Press.

Felsenstein, J. 1993. PHYLIP (phylogeny inference package) version 3.5 c. *Department of Genetics, University of Washington, Seattle*, vol. 1118.

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 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.

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.

Gilij, Filippo Salvatore 2001. Saggio di storia americana, osia storia naturale, ciuile, e sacra de regni, e delle provincie spagnuole di terra-ferma nell’america meridional/descrita dall’abate filippo salvadore gilij.-roma: per luigi perego erede salvioni..., 1780-1784. *Textos clásicos sobre la historia de venezuela:[recopilación de libros digitalizados]*, 11. MAPFRE.

Goddard, Cliff 2001. Lexico-semantic universals: A critical overview. *Linguistic Typology*, pp. 1–65.

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.

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.

Guy, Jacques B. M. 1994. An algorithm for identifying cognates in bilingual wordlists and its applicability to machine translation. *Journal of Quantitative Linguistics* 1 (1): 35–42.

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 2013. Basic vocabulary comparison in South American languages. Pieter Muysken and Loretta O'Connor (eds), *Native languages of South America: Origins, development, typology*, 126–151. Cambridge: Cambridge University Press.

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.

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.

Heeringa, Wilbert Jan 2004. Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. diss., University Library Groningen.

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. Continuum International Publishing Group.

Hock, Hans Heinrich 2010. Typology and universals. Silvia Luraghi and Vít Bubeník (eds), *Continuum companion to historical linguistics*, 59–69. Continuum International Publishing Group.

Hock, Hans Henrich 1991. *Principles of historical linguistics*. 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. 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. 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.

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.

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*. UK: Cambridge University Press.

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 2014. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. Forthcoming.

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.

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 2005. Phonetic comparison algorithms. *Transactions of the Philological Society* 103 (2): 243–260.

Kessler, Brett 2008. The mathematical assessment of long-range linguistic relationships. *Language and Linguistics Compass* 2 (5): 821–839.

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 2005. Europarl: A parallel corpus for statistical machine translation. *Mt summit*, Volume 5.

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.

Koehn, Philipp and Kevin Knight 2002. Learning a translation lexicon from monolingual corpora. *Proceedings of the ACL-02 workshop on unsupervised lexical acquisition*, Volume 9, 9–16. 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 2005a. Cognates and word alignment in bitexts. *Proceedings of the tenth machine translation summit (mt summit x)*, 305–312.

Kondrak, Grzegorz 2005b. N-gram similarity and distance. *String processing and information retrieval*, 115–126. Springer.

Kondrak, Grzegorz 2009. Identification of cognates and recurrent sound correspondences in word lists. *Traitement Automatique des Langues et Langues Anciennes* 50 (2): 201–235 (October).

Kondrak, Grzegorz and Bonnie Dorr 2006. Automatic identification of confusable drug names. *Artificial Intelligence in Medicine* 36 (1): 29–42.

Kondrak, Grzegorz, Daniel Marcu and Kevin Knight 2003. Cognates can improve statistical translation models. *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology: companion volume of the proceedings of HLT-NAACL 2003–short papers*, Volume 2, 46–48. Association for Computational Linguistics.

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.

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. Dallas, TX: SIL International. Online version: `http://www.ethnologue.com`.

Lewis, Paul M. (ed.) 2009. *Ethnologue: Languages of the world.* Sixteenth. 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.

Lin, Dekang 1998. An information-theoretic definition of similarity. *Proceedings of the 15th international conference on machine learning*, Volume 1, 296–304.

Lin, Jimmy and Chris Dyer 2010. Data-intensive text processing with MapReduce. *Synthesis Lectures on Human Language Technologies* 3 (1): 1–177.

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.

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. Continuum International Publishing Group.

Maddieson, Ian and Kristin Precoda 1990. Updating UPSID. *UCLA working papers in phonetics*, Volume 74, 104–111. Department of Linguistics, UCLA.

Mallory, James P. 1989. *In search of the Indo-Europeans: language, archaeology and myth.* Volume 186. London: Thames and Hudson.

Marzal, Andres and Enrique Vidal 1993. Computation of normalized edit distance and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 15 (9): 926–932.

Masica, Colin P. 1993. *The Indo-Aryan languages.* Cambridge Language Surveys. Cambridge: Cambridge University Press.

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.

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.

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.

Nelson-Sathi, Shijulal, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin and Tal Dagan 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B: Biological Sciences* 278 (1713): 1794–1803.

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 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 (ed.) 2012. *Electronic Grammaticography*. Honolulu, Hawaiı́: University of Hawaiı́.

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.

Parkvall, Mikael 2009. *Sveriges språk-vem talar vad och var? (The languages of Sweden. Who speaks what and where?).* Institutionen för lingvistik, Stockholms universitet.

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.

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.

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.

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 and Lars Borin 2011. Estimating language relationships from a parallel corpus. A study of the Europarl corpus. *NEALT Proceedings Series (NODALIDA 2011 Conference Proceedings)*, Volume 11, 161–167.

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 2014. Comparison of string similarity measures for automated language classification. Under review.

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.

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.

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 University Press.

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.

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.

Schleicher, August 1853. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Sprachwissenschaft und Literatur* 1853: 786–787.

Schmidt, Johannes 1872. *Die Verwantschaftsverhältnisse der indogermanischen Sprachen*. Weimar: Böhlau.

Serva, Maurizio and Filippo Petroni 2008. Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)* 81: 68005.

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.

Singh, Anil Kumar, Harshit Surana and Karthik Gali 2007. More accurate fuzzy text search for languages using Abugida scripts. *Proceedings of ACM SIGIR workshop on improving web retrieval for non-english queries*. Amsterdam, Netherlands.

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.

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.

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 SBM. The role of language evolution in digital archives. Submitted (SBM).

Tate, Robert F. 1954. Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of mathematical statistics* 25 (3): 603–607.

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.

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.

Wang, William S-Y. and James W. Minett 2005. Vertical and horizontal transmission in language evolution. *Transactions of the Philological Society* 103 (2): 121–146.

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 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. Continuum International Publishing Group.

Wichmann, Søren 2010b. Neolithic linguistics. Forthcoming.

Wichmann, Søren 2013a. Genealogical classification. Oxford Bibliographies Online: Linguistics.

Wichmann, Søren 2013b. 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, 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.

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*. Wordsworth Editions.

A Supplementary information to phonotactic diversity

This chapter contains the supplementary information (tables and figures) referred in paper IV.

A.1 Data

Table A.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 A.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.

A.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 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 omitting influences the error value.
- A case plot of the Cook's statistic.

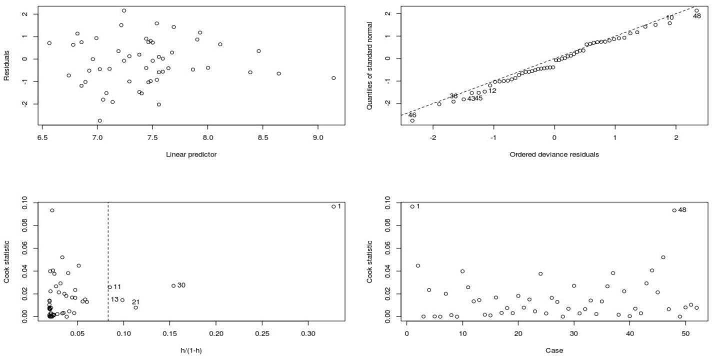


Figure A.1: Diagnostic plots for 1-grams

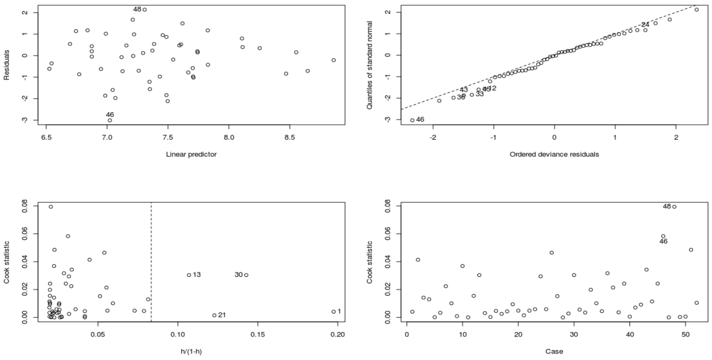


Figure A.2: Diagnostic plots for 2-grams

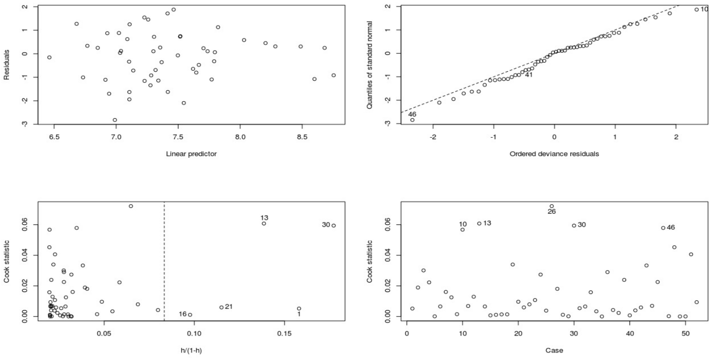


Figure A.3: Diagnostic plots for 3-grams

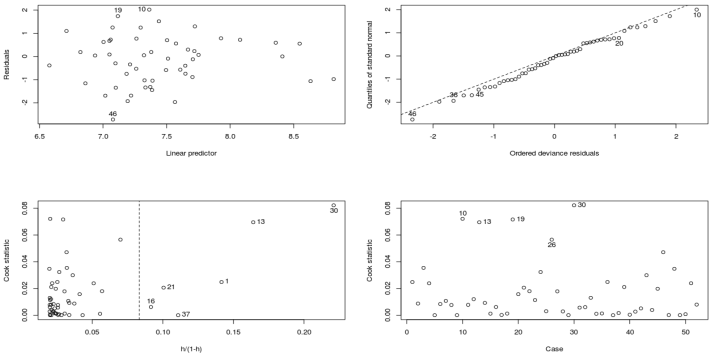


Figure A.4: Diagnostic plots for 4-grams

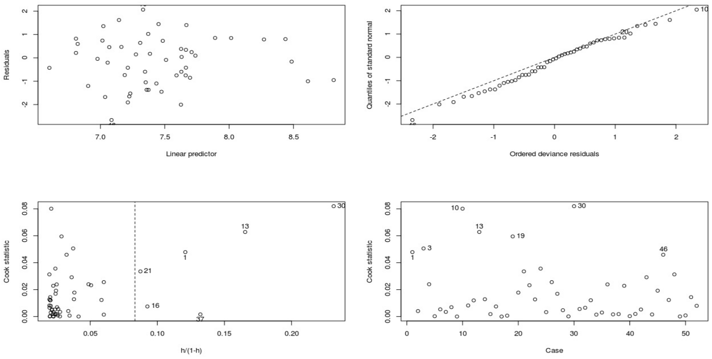


Figure A.5: Diagnostic plots for 5-grams

A.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 A.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 A.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 A.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 A.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 A.6: Dates for language groups of South America

B Appendix to evaluation of string similarity measures

This chapter contains the supplementary information to the tables referred in paper V.

B.1 Results for string similarity

The tables B.1, B.2, and B.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 B.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 B.2: Dist for families and measures above average

Family	LDND	LCSD	LDN	LCS	PREFIXD	PREFIX	DICED	DICE	JCD	JCDD	TRID[58]
NDe	0.5761	0.5963	0.5556	0.5804	0.5006	0.4749	0.4417	0.4372	0.4089	0.412	0.2841
Bos											
NC	0.4569	0.4437	0.4545	0.4398	0.3384	0.3349	0.3833	0.3893	0.3538	0.3485	0.2925
Hok	0.8054	0.8047	0.8048	0.8124	0.6834	0.6715	0.7987	0.8032	0.7629	0.7592	0.5457
Pan											
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	0.6385	0.6413	0.5763	0.5759	0.6056	0.538	0.5816	0.5176	0.5734	0.5732	0.5147
Que											
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	0.2971	0.2718	0.3092	0.3023	0.2926	0.3063	0.2867	0.257	0.2618	0.2672	0.2487
Tuc											
Ura	0.4442	0.4356	0.6275	0.6184	0.4116	0.6104	0.2806	0.539	0.399	0.3951	0.1021
Arw											
May											
LP	0.41	0.4279	0.4492	0.4748	0.3864	0.4184	0.3323	0.336	0.3157	0.3093	0.1848
OM	0.8095	0.817	0.7996	0.7988	0.7857	0.7852	0.7261	0.7282	0.6941	0.6921	0.6033
Car											
MZ											
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 B.3: RW for families and measures above average.

B.2 Plots for length normalization

The following hexagonal binned plots show the relation between LD and length for word pairs from the Dravidian language family.

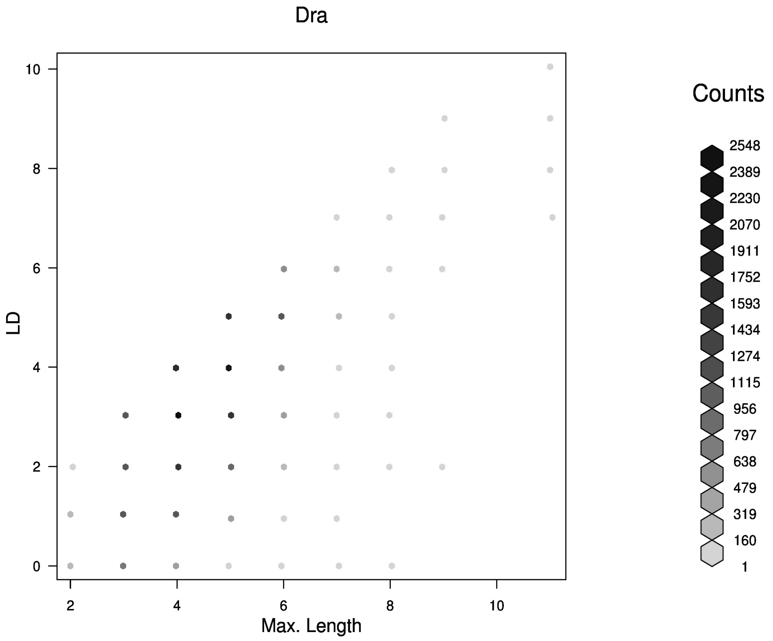


Figure B.1: Hexagonally binned plot of same meaning-meaning LD and maximum length.

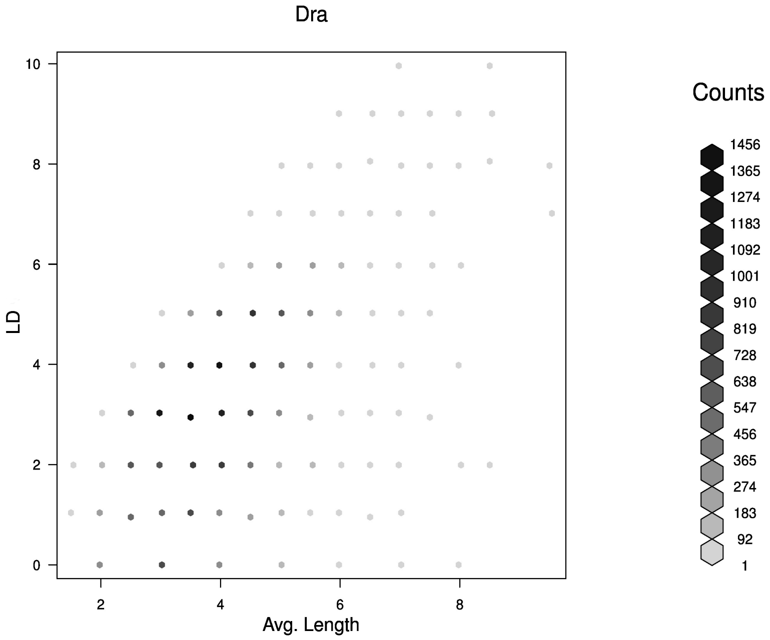


Figure B.2: Hexagonally binned plot of same meaning-meaning LD and average length.

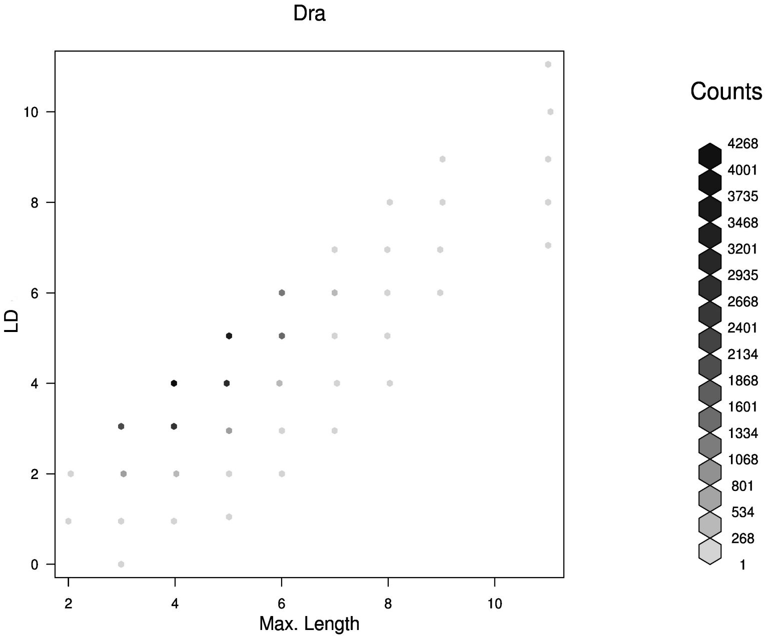


Figure B.3: Hexagonally binned plot different meaning-meaning LD and maximum length.

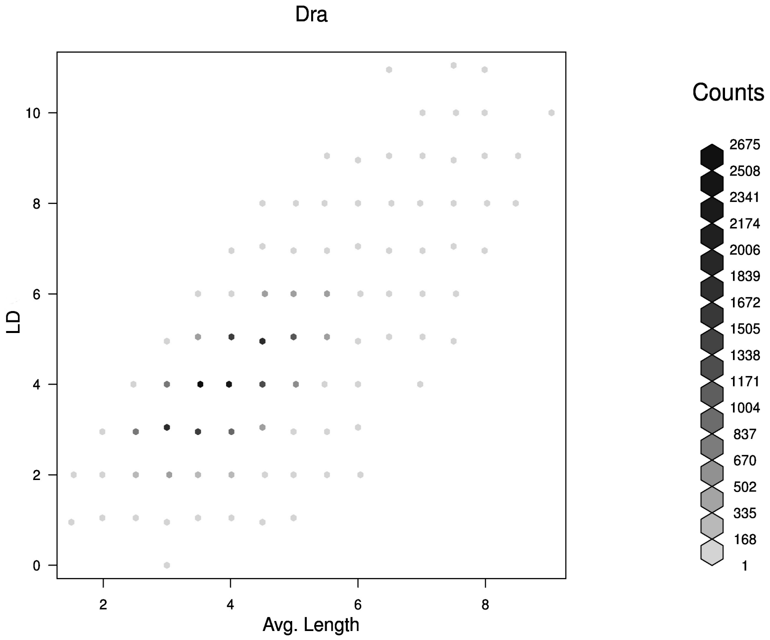


Figure B.4: Hexagonally binned plot different meaning-meaning LD and average length.