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$\bar{\text{V}}$- 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,
Sī̄ 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 *sī̄* (related to German *sind*) and English *be* are descendants from different PIE roots *h_1*es*- and *b^h*u*h_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 *s̥ṛngavera* '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 semi-

supervised 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, t̲, l, l̤, z̤, 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-* ~ *θreis* 'three'
- C {+Voicing, +Aspiration} ~ C {+Voicing, -Aspiration}: *madhya-* ~ *midjis* 'middle'
- C {+Voicing, -Aspiration} ~ C {-Voicing, -Aspiration}: *daśa-* ~ *taihun* 'ten'

However, there were exceptions to this law: whenever the voiceless plosive did not occur in the word-initial position or did not have an accent in the previous syllable, the voiceless plosive became voiced. This is known as *Verner's law*. Some examples of this law are: Sanskrit *pitár* ~ Old English *faedar* 'father', Sanskrit *(va)vrtimá* ~ Old English *wurdon* 'to turn'.

The 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 $*d^hug^h+t^has \overset{BL}{\rightarrow} *d^hug^hd^has$

$\overset{GL}{\rightarrow}$ *dug^hd^has $\overset{deaspiration}{\rightarrow}$ $dugd^has$. 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^hā-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^hag^h-, but 3 sg. s-aor. *á-d^hāk* < **-dhāk-s-t*.

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

Another well-known sound change in 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 *ke̱ṭu > 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 – *enḍa*, *nela*, and *cukka* – became less frequent or were relegated to specific contexts. Brahui, a Dravidian language surrounded by Indo-Aryan languages, also borrowed quite a large number of basic vocabulary items.

The second evidence for language relationship comes from sound correspondences. Sound correspondences should be recurrent and not sporadic. The

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

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

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

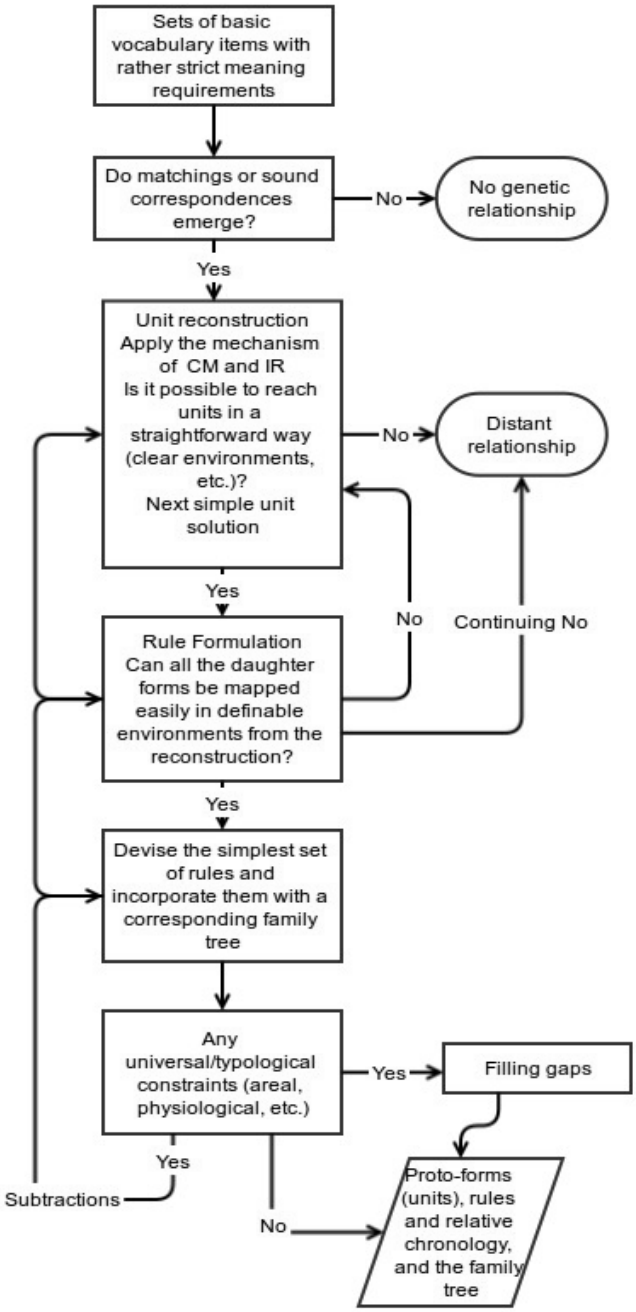


Figure 2.1: Flowchart of the reconstruction procedure (Anttila 1989: 347). CM and IR stand for the comparative method and internal reconstruction.

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 protoforms 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 protoform **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^ha- ~ 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

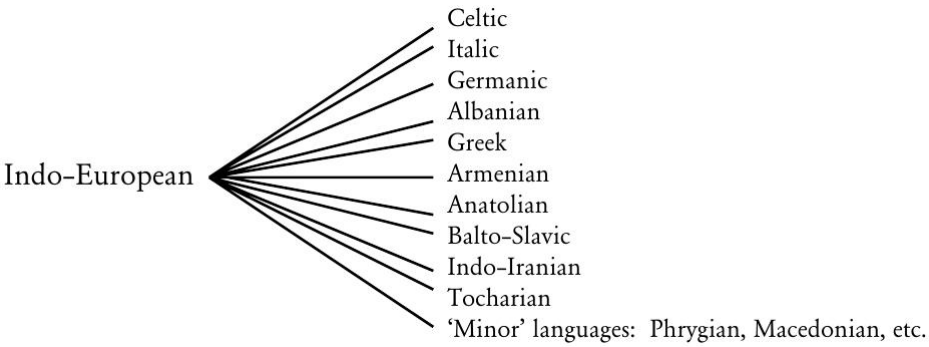


Figure 2.2: Higher-order tree of IE family from Garrett (1999).

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 $s > h > \emptyset$ in Gondi dialects by Krishnamurti (1998) is an example where the original Proto-Dravidian $*c > *s$ in the word-initial, pre-vocalic position completed the sound change in

South Dravidian languages. This sound change is succeeded by $*s > *h > \emptyset$ and is completed in South Dravidian I and Telugu. The same sound change is still ongoing in some Gondi dialects and the completion of the sound change marks the dialectal boundary in Gondi.

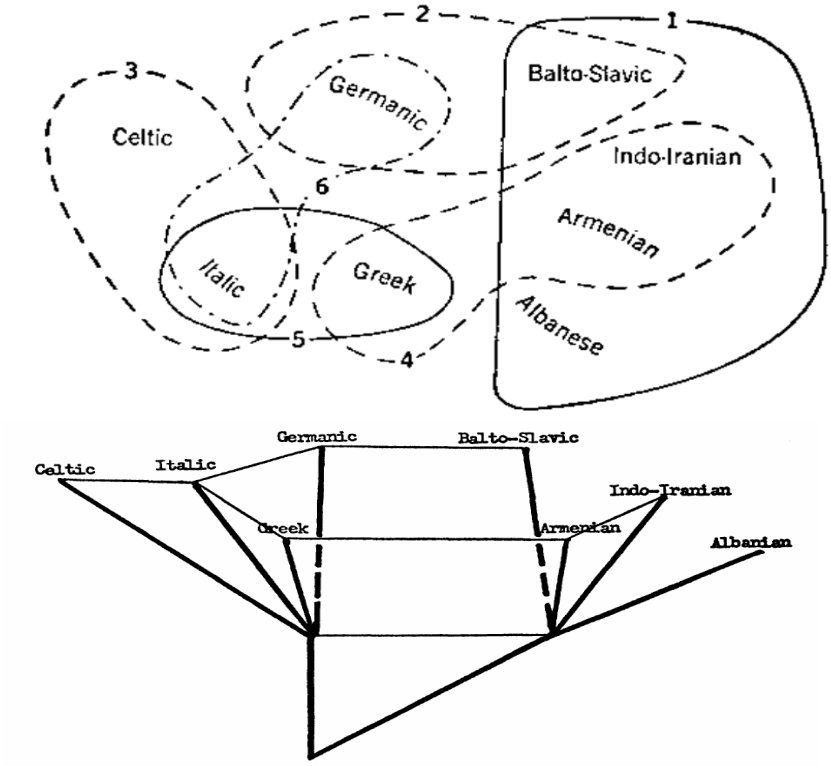
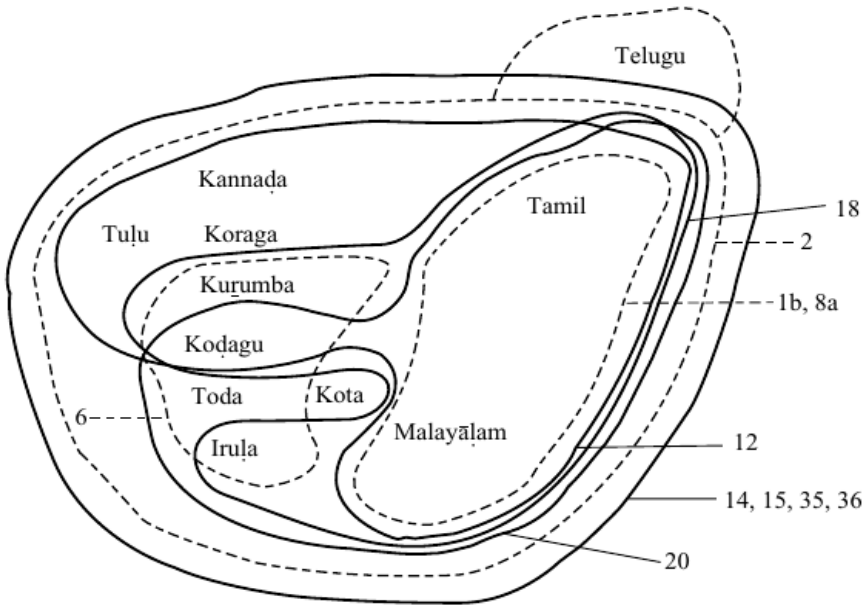


Figure 2.3: Indo-European isoglosses (Bloomfield 1935: 316) and the corresponding tree-envelope representation from Southworth (1964). The numbers in isogloss figure correspond to the following features. **1.** Sibilants for velars in certain forms. **2.** Case-endings with [m] for [b^h]. **3.** Passive-voice 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-aṉṯ*, **iw-aṉṯ* 'he'
F18. Loss of -V*n* 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

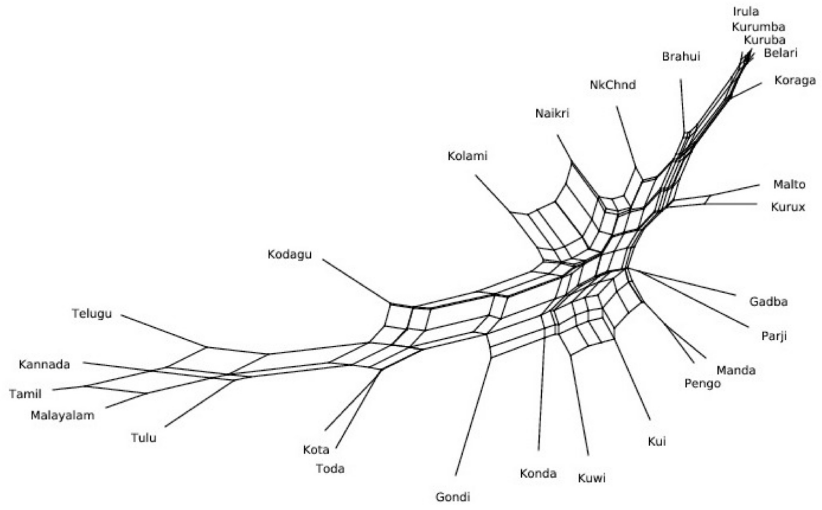


Figure 2.5: A network diagram of 28 Dravidian languages based on grammatical and phonological features (Rama and Kolachina 2013).

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 prob-

abilities (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 classifica-

tions 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).

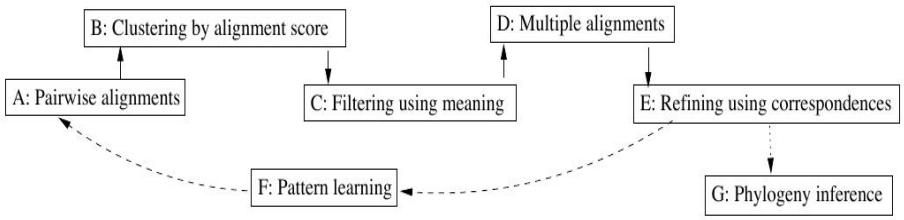


Figure 2.6: A pipeline for a language classification system.

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:

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

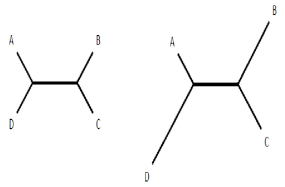


Figure 2.7: Additive trees for a quartet

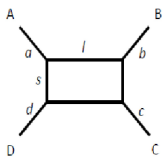


Figure 2.8: Reticulate quartet

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 multi-

gram 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.