Cognate Identification and Phylogenetic Inference : Search for a Better Past

Abhaya Agarwal
Jason Adams

November 21, 2007

1 Introduction

Historical linguistics studies the relationships between languages as they change over time. Occasionally, speakers of a language will split into separate groups for any number of reasons and become isolated from each other. When this happens, the language they shared begins to diverge. These divergences are typically phonological, syntactic, and semantic in nature. These changes in each child language occur in different but systematic ways. Studying this phenomenon is at the heart of diachronic linguistics.

In this paper, we will examine the role of two main applications of computational methods to historical linguistics. The first is the identification of cognates. Identifying cognates is the first step in the comparative method, the primary technique used by historical linguists to determine the relatedness of languages. The second area is phylogenetic inference, a method of automatically reconstructing the genetic relationships between languages and language families. We will examine computational and statistical approaches to these topics that have been taken in the literature to present a picture of the state of the art, concluding each section with a discussion of the future work to be done in these areas.

In Section 2, we provide background information on historical linguistics that will inform the rest of this paper. Cognate identification and phonetic similarity measures are discussed in Section 3. Cognate identification forms the basis for reconstructing the evolutionary history of languages. We will briefly discuss some early work on this problem in Section 4 and then go on to discuss more recent methods of phylogenetic inference in Section 5. Finally, we present our conclusions about the application of computational methods to historical linguistics in Section 6.

2 Historical Linguistics

The field of historical linguistics (also called diachronic linguistics) is concerned primarily with the changes languages undergo over time and determining the relatedness of languages. These changes usually take the form of phonological, syntactic and semantic changes.

One might imagine the scenario of two groups of people speaking the same langauge separating. Phonological innovations and transformations occur, causing the emergence of dialects and regional accents. An important observation in historical linguistics is that phonological changes that occur in this manner are almost always regular. The same sound in the same context will change the same across the entire language. Given enough time, dialects diverge and ultimately lead to two different languages, where the speakers are no longer able to understand each other.

Words in daughter languages sharing a common ancestor are known as cognates. In the context of historical linguistics, this does not include words **borrowed** from other languages. Borrowed words are words introduced directly from one language to another. Borrowed words do not undergo the sort of phonological changes that occur over time and so differ from cognates. Cognates contain helpful information for historical linguistics because they contain the regular sound changes that led to the formation of the divergent languages. For example, the English word *beaver* and the German word *Biber* are cognates, both descending from Proto-Germanic **bebru* and Proto-Indo-European **bher* (Köbler, 1980). A systematic phonological change in the history of English caused the second /b/ to become realized as /v/ (denoted by the letter *f* in this context) in Old English *beofor*. Because cognates are so useful in helping isolate diachronic sound changes (phonological changes that occur over time), the task of cognate identification is a central element in the study of diachronic linguistics (Kondrak, 2003b). Automatic methods for handling cognate identification are described in more detail in Section 3.

The primary method historical linguists use to determine whether two languages are related is known as the comparative method (Hoenigswald, 1960). This method begins by compiling a list of probable cognates. Once a set of cognates has been identified, historical linguists use the comparative method to infer the series of phonological changes that led to the divergence in the two daughter languages under comparison. These phonological changes are realized in cognate lists as **sound correspondences**. A sound correspondence is a mapping of a phoneme or series of phonemes in one language to a phoneme or set of phonemes in another related language. In German-English example above, the modern English sound /v/ corresponds to the modern German sound /b/ in this context. Once sound correspondences have been identified, the next step in the comparative method is to infer the phonological changes that led to the correspondences and use these to reconstruct the word forms in proto-language.

Many of the languages seen in the world today have evolved through the process described above, like Latin evolved into modern romance languages like Spanish, French and Italian. This suggests that languages can be grouped into

families, sub families and so on based on common ancestory. Indo-European is a example of one such family. Phylogenetic inference (Section 5) is concerned with identifying these groupings based on the features observed in these languages.

Greenburg proposed a method called mass lexical comparison or multilateral comparison for finding genetic relationships that cannot be captured by the comparative method. Instead of looking for recurrent sound correspondances, Greenburg looked at the surface similarity of the words in various languages and used them to propose long distance genetic relationships. This methods has been sharply criticized by many linguists for its lack of rigour and is not in wide spread use. However as we shall see in Section 5.6, it looks attractive from a computational perspective.

Let us start by looking at the computational techniques for cognate identification in the next section.

3 Cognate Identification

Identifying sound correspondences and cognates can be a time-consuming and laborious process, requiring the expert knowledge of a linguist. There are still many languages that have received little attention due to the amount of effort involved. Finding automatic methods for performing or bootstrapping these processes would be a great benefit to historical linguists and has been a major motivation for research on cognate identification. Achieving good performance on automatic cognate identification can also benefit machine translation when dealing with two languages that share a certain quantity of cognates, as cognates are usually translations and serve as anchors when aligning bitexts.

The datasets used in the cognate identification literature are as varied as the approaches. There is no single dataset that dominates the field. Most papers use datasets that are relatively small and are typically dictionaries. A few examples of datasets include a dictionary of Algonquian languages produced by Hewson (Hewson, 1993; Kondrak, 2001, 2002b, 2003b, 2004), a set of 82 cognate pairs derived from Swadesh lists (Covington, 1996; Kondrak, 2000, 2003a), a small German-English dictionary[1] (Mulloni and Pekar, 2006), and the Wordgumbo online English-Polish and English-Russian dictionaries[2] (Ellison, 2007). Swadesh lists were created by Morris Swadesh for comparing languages. A Swadesh list is a set of 100 or 200 words for a particular language family that represent the most common and most useful words (Swadesh, 1955a). These lists often serve as datasets for cognate identification. The Comparative Indo-European Data Corpus (Dyen et al., 1992a), used by (Mackay and Kondrak, 2005) and (Kondrak and Sherif, 2006) consists of Swadesh-style lists of 200 words for 95 languages of the Indo-European language family.

Evaluation of cognate identification methods typically uses measures of precision, recall, and accuracy. Less often are F1 scores reported as well, but precision and accuracy are the predominant measures. Precision (equation 1)

[1] Available http://www.june29.com/IDP.
[2] Available http://www.wordgumbo.com.

and recall (equation 2) are calculated as they are for information retrieval:

$$P = \frac{|\text{correct cognates identified}|}{|\text{cognates identified}|} \tag{1}$$

and

$$R = \frac{|\text{correct cognates identified}|}{|\text{cognates in test set}|}. \tag{2}$$

Accuracy is the ratio of correct cognates to the total number of cognates according to a gold standard list. The F1 measure, when reported, is the harmonic mean of precision and recall:

$$F1 = \frac{2PR}{P+R} \tag{3}$$

Identifying correspondences and cognates is a difficult task because several sound changes may have occurred between two languages, obscuring their relatedness. The inherent difficulty in identifying correspondences and cognates has motivated the development of several computational tools and methods for facilitating the process. Another factor driving the automatic identification of cognates is machine translation, specifically learning translation correspondences. In the case of historical linguistics, most approaches deal with phonetic similarity and sound correspondences. Phonetic similarity attempts to measure how similar phonemes in two languages are. Measures of dialect and language distance rely heavily on phonetic similarity. These disparate needs have motivated two main ways of handling cognate identification: **orthographic** and **phonetic**. Orthographic methods focus on the characters used in the writing system and make the assumption that characters correspond to consistent sounds in their respective languages. Phonetic methods rely on phonetic transcriptions of the languages and look at phonetic similarity amongst other things.

Approaches to cognate identification use both manually constructed schema and empirical methods (not necessarily together). Orthographic methods tend to rely more on empirical methods, whereas several of the earlier phonetic methods rely more on manually constructed schema. The most recent approaches lean more strongly towards empirical methods. A recent evaluation by Kondrak and Sherif (Kondrak and Sherif, 2006) hints that empirical methods may be performing best, but it is still an open question as to which will prevail.

In computational approaches, the strict definition of cognate from historical linguistics is usually abandoned. In historical linguistics, a cognate is a word in two related languages that has a single parent in an ancestor language. Computational approaches typically discard the constraint that the two words descend from the parent. Borrowed words are words that enter a language directly from another language. Borrowings can take on many forms, sometimes appearing in the target language untranslated from the original. English examples include *bagel* and *avatar* from Yiddish and Sanskrit, respectively. Other times they are borrowed in translated form, such as *worldview* from the German *Weltanschauung*. In the case of the former, most cognate identification

algorithms do not distinguish between such borrowings and cognates. In the case of the latter, most algorithms would fail to recognize the two as cognates. Hereafter, the term cognate will refer to the looser computational definition and **strict cognate** to the definition from historical linguistics (the looser definition is sometimes referred to as an **orthographic cognate**).

The remainder of this section is organized by first looking at orthographic approaches in Section 3.1. Often these approaches are coupled with a particular application, such as bitext alignment. The method and the application will be discussed in tandem. In Section 3.2 we will discuss issues of determining phonetic similarity. There is no agreement of the exact determination of phonetic similarity between phonemes in different languages so a variety of approaches will be discussed. Phonetic similarity plays a key role in phonetic cognate identification methods, which will be discussed in Section 3.3. In cases where the techniques are blended, we will include them with the phonetic approaches. Throughout these sections we will look at the algorithms and formalisms used, the various datasets examined, and the evaluation techniques in an attempt to compare the various approaches. We conclude the section in Section 3.4 with a discussion of future work and the approaches covered.

3.1 Orthographic Methods

The simplest orthographic approach is full string matching. If two words are identical across languages they are hypothesized to be cognates. This approach is naïve as it may find **false friends** (*faux amis*). False friends are words in two languages that have the same orthographic or phonological realization (depending on the data being used) but have very different meanings or origins. For example, *Billion* in German means *trillion* instead of the English *billion* that it resembles. Even if the technique does not care about finding strict cognates, this still poses problems for most applications. In machine translation, it is necessary to align parallel corpora at the sentence and word level. A number of approaches have incorporated cognate identification into this task (Simard et al., 1993; Church, 1993; Melamed, 1999). If false friends are always aligned the proposed translations could suffer.

3.1.1 String Matching

One of the earliest approaches to orthographic cognate identification was string matching (Simard et al., 1993). In this paper, Simard et al. approached the task of aligning sentences in bilingual corpora by looking at a measure of **cognateness** in each possible sentence. The measure of cognateness (or cognacy) is an attempt to quantify the extent to which two words are cognates. Potential cognates are found by looking at the first four characters in a word. If they match, they are hypothesized to be cognates. The cognateness γ of two candidates for alignment (one from each language) is computed as

$$\gamma = \frac{c}{(n+m)/2}, \tag{4}$$

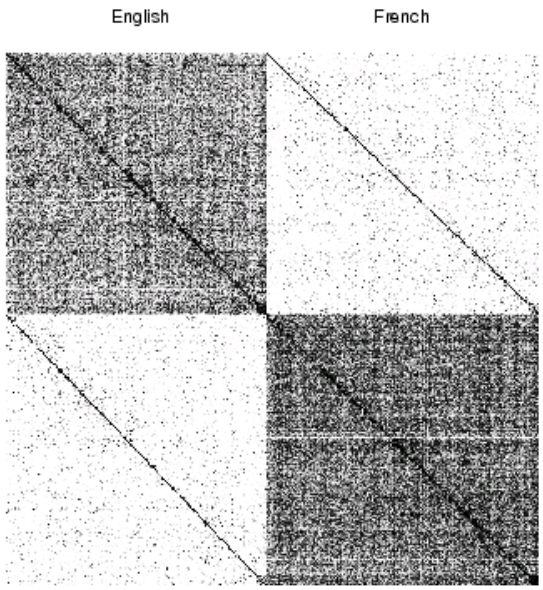

Figure 1: Dot-plot of three years of Hansards (37 million words) (Church, 1993).

where there are c pairs of cognates in the bilingual segments of length n and m. This model relies on the assumption that a translation correlates with cognateness, so that there will be more cognates in a translation pair than in a randomly selected pair. Simard et al. hand-aligned a portion of the Hansard corpus and found that the average cognateness of translations were indeed significantly higher than the average cognateness of random pairs.

Kenneth Church took a similar approach to Simard et al. with the Char_align system (Church, 1993). Like Simard et al., he used the first four characters of potential word pairs to determine cognates. His approach differed in that it no longer used a measure of cognateness but a sort of histogram he calls a dot-plot (see Figure 1). At each possible alignment in the texts, if the two words are cognates a dot is placed on the graph. Additional signal processing techniques are used to remove noise and the result is a line that represents the best alignment. Church employed a sub-optimal heuristic search with forward pruning to find paths. Each path was scored by the sum of its weights and the one with the highest score was chosen. Church admits the procedure is *ad hoc*, but the work is important because it demonstrates that statistical techniques can be used to find cognates reliable enough to align a bitext.

3.1.2 Dice's Coefficient

Several approaches use Dice's coefficient as a comparison between two strings to determine if they are cognates. Dice's coefficient is typically calculated on two sets of items to determine similarity between the two sets. Let f_x and f_y be the occurrences of items in a collection of aligned units and $f_{x,y}$ the co-occurrences of items in the aligned set (Tiedemann, 1999). Dice's coefficient is then calculated as

$$Dice(x, y) = \frac{2f_{x,y}}{f_x + f_y}. \tag{5}$$

In Section 3.1.4, we further describe Tiedemann's approach, which makes heavy use of the Dice coefficient in conjunction with dynamic programming.

Brew and McKelvie use one such approach for the task of bilingual lexicon building as a tool for lexicographers (Brew and McKelvie, 1996). They use a variant of Dice's coefficient originally used by McEnery and Oakes for the task of sentence and word alignment (McEnery and Oakes, 1996). Brew and McKelvie looked at character bigrams for hypothesis word pairs. Dice's coefficient was then calculated in five different ways. Besides looking at the standard Dice formulation as in equation 5, they changed how bigrams were found and weighted. Extended bigrams were found by taking sequences of three letters and removing the middle letter. The weighted Dice measure added weights to the counts in the Dice equation in inverse proportion to their frequency. The weight of a bigram was calculated as

$$weight(bigram_i) = \frac{N_{tokens} + N_{types}}{freq(bigram_i) + 1}, \tag{6}$$

where N_{tokens} is the number of tokens seen in both of the bilingual corpora and N_{types} is the number of distinct word types in those corpora. Another variant combined extended bigrams with the weighted Dice formula. Another variant penalized matches from different parts of the words. Lastly, they used longest common subsequence (see Section 3.1.3).

Their methods were tested on the French and English sections of the MLCC Multilingual Corpus[3], a collection of parliamentary questions taken from 1992-1994. Brew and McKelvie were trying to create a tool for lexicographers that returned possible translation pairs, potential cognates, as well as false friends. One goal of the tool was to return results with variable precision and recall. High precision items were cognates while high recall items were intended to consist of many false friends. The variant of the Dice's coefficient that penalized matches from different parts of the word returned the best precision. However, none of the measures used were able to reliably detect false friends. High recall items included many false translations and artifacts that would have been rejected by humans.

[3] Available http://www.elda.org/catalogue/en/text/W0023.html.

3.1.3 Longest Common Subsequence Ratio

The Longest Common Subsequence Ratio (LCSR), as the name suggests, is an attempt to measure partial matches between two words by looking at the length of matching substrings. The LCSR is calculated as

$$LCSR(x, y) = \frac{\|LCS(x, y)\|}{max(\|x\|, \|y\|)}, \tag{7}$$

where the $LCS(x, y)$ is the longest common subsequence of segments x and y.

One application of LCSR to the task of aligning bitexts builds on the work by Simard et al was done by Melamed (1999). Melamed recognized the two problems in that approach of finding cognates as being false positives and false negatives. In Simard et al. (1993), a cognate was determined only by matching the first four characters. True cognates that differed in that substring were discarded, resulting in false negatives. Likewise, false positives occurred whenever common prefixes led to matches. For example, English and French both share the prefix *con*. This leads to false positives such as *conseil* and *conservative*. His solution was to use LCSR with a threshold that was dependent on the two languages under consideration. This threshold was related to the level of relatedness between the two languages. Melamed acknowledged this approach to finding cognates could be extended to the phonetic level, given phonetic transcriptions of the source texts.

With the set of cognates identified, his algorithm uses those as points of reference for determining alignments. While he does not evaluate the accuracy of cognate identification directly, his results show improvements in accuracy for the task of bitext mapping over previous work on the Hansard corpus. Melamed reports improvement using LCSR over the naïve string matching of Simard et al. Later work by Kondrak looks more closely at LCSR and compares it to phonetic methods (see Section 3.3.1).

3.1.4 Weighted String Similarity Measures

Jörg Tiedemann developed three weighted string similarity measures that can be used for cognate identification (Tiedemann, 1999). His approach is at the furthest end of the orthographic spectrum, drawing in information about vowels and consonant sequences. His approaches seek to find a matching function m based on different factors. He evaluated each approach using the PLUG Corpus of aligned Swedish-English technical texts.

The first approach is called *VCchar* and scores pairs of vowels and consonants higher that co-occur in the reference lexicon more often. Vowels and consonants at similar positions in the word pairs are counted and the Dice coefficient scores them. The resulting list is sorted in order of the Dice coefficient and the best scores are examined. Swedish and English return mostly identical pairs of characters for the top scores. A threshold value is then optimized on the development data to achieve greater precision.

The second approach Tiedemann uses is *VCseq*, which looks at matching sequences of vowels and consonants (rather than single characters as in VCchar). The method of computing the m-function is similar to the first approach, but words are segmented differently. Word pairs are split into sequences of vowels followed by sequences of consonants and vice versa. Characters that interrupt sequences appear in neither. Sequences at identical estimated positions are scored as matches and the ones with the highest Dice values are examined. Again a threshold value was applied to achieve higher precision. This method finds longer sequences of potential matches.

The third approach is called *NMmap*. LCS is used in conjunction with dynamic programming to find non-matching parts of the two strings. Each pair of non-matching strings $[x, y]$ are given a weight as the ratio of the frequency of $[x, y]$ and the frequency of x. This approach attempts to find systematic differences in orthographic systems. The highest scoring example is the correspondence between Swedish *ska* and English *c*. This difference manifests in word pairs such as *asymmetriska* and *asymmetric*. Of the three methods, Tiedemann found that the NMmap was the best if languages with a fairly common character set were used.

Weighted string similarity measures have also been used in the discovery of bitext (Smith, 2001). Bitext discovery is the task of automatically finding document pairs that are mutual translations. Cognates were found to be helpful for detecting translation equivalence since it was not necessary that they had been seen in the data previously. Smith modified Tiedemann's approach, which trained on a list of known cognate pairs, by instead training using a statistical translation model. Translation models require aligned bitext for training and provide probabilities of translational equivalence for bilingual word pairs. These probabilities were incorporated into the matching function (for details, please see (Smith, 2001)). Smith presented results showing that this approach could reliably classify whether two texts were bitexts.

3.1.5 Multigram Alignment

A common approach in phonetic methods is to align pairs of words based so that corresponding phonemes are matched (see Section 3.2). This has the advantage of removing the orthography from the equation by looking at the actual phonemes involved, which are closer to the ground truth of spoken language. However, the phonetic approach has the disadvantage of scarcity of data or reliance on automatic methods (noisy) for producing phonetic transcriptions. Some recent work has been done on producing alignments using orthography alone, which has the advantage of large amounts of data (Cysouw and Jung, 2007). The assumption underlying this approach is that sound correspondences will still emerge from the data and allow reliable alignments to be formed, even with non-identical orthographies.

Cysouw and Jung describe an iterative process by which they match **multigrams** (sequences of characters of varying length) using a variant of Levenshtein distance. Levenshtein distance is one way of measuring edit distance that works

on strings of differing lengths and counts the number of insertions, deletions, and substitutions necessary to transform one string into another. Cysouw and Jung extend this by allowing for mappings of variable length (whereas Levenshtein compares only one character at a time) and by assigning a cost between 0 and 1 to each operation. The cost for each multigram is found first by counting the co-occurrences of the multi-grams in language word lists for each language. The Dice coefficient is calculated for each possible subsequence as a cost function. To cut down on computation time, they limit the size of multi-grams to four characters. Multi-gram costs were length-normalized so that all values fell in the 0 to 1 range.

Cysouw and Jung evaluate their system using data from the Intercontinental Dictionary Series (IDS) database.[4] They extracted about 900 word pairs from English-French, English-Hunzib (a Caucasian language), and Spanish-Portuguese. They found that reliable alignments were still possible even without phonetic transcriptions and these could be used to find cognates.

3.2 Phonetic Similarity

The quantification of phonetic similarity is an important component in diachronic and synchronic phonology (Kondrak, 2003a). However, computing phonetic similarity is not always straightforward. If we used edit distance to measure similarity, relatively similar phonemes /d/ and /t/ would be given the same weight as /d/ and /a/, which are quite different from the standpoint of human intuition. Care must be taken to produce phonetic similarity measures that match human intuition and linguistic realization more closely.

One early application of evaluating phonetic similarity was for the task of aligning suspected cognates for historical comparison (Covington, 1996). Such word alignments serve as the first step in applying the comparative method, which seeks to establish historical relationships between languages. Alignments are found by comparing surface forms of phonemes in a method that is meant to mirror a linguist's first look at unfamiliar data. Except in the trivial case of exact matches, all alignments are attempts at inexact string matching. To produce alignments, the aligner moves through the two strings performing either a skip or a match. A cost is assigned to each. For each possible alignment, the alignment is scored and an n-best list returned. Computation time is decreased by computing the score as alignments are searched, giving up on a possible alignment as soon as it exceeds the best value thus far. In this way, the aligner maximizes the phonetic similarity scores of characters in a word-pair (by minimizing the cost).

In Covington's first formulation, the actual phonemes do not play a large role. He argues that alignment looks more at the placement of a sound in a word rather than the paradigm for the sound in the language. He goes on to make the point that by adding feature-based phonology, the system could improve. Phonemes that are produced at closer positions in the mouth should

[4] Available http://www.eva.mpg.de/lingua/files/ids.html.

be considered more similar. Taking this into account would have corrected some of the errors his system made. He evaluated his system on 82 pairs of cognates from several different languages drawn from the Swadesh lists of Ringe (Ringe, 1992). No automatically derived evaluation score was given but he reproduces lists of output for various language pairs from manual inspection. Results were better for languages that are more closely related and worse for those which were not.

3.2.1 ALINE

Kondrak has contributed a large body of work to the tasks of computing phonetic similarity and identifying cognates (Kondrak, 2000, 2001, 2002a,b, 2003a,b; Kondrak et al., 2003; Kondrak, 2004; Inkpen et al., 2005; Kondrak, 2005; Mackay and Kondrak, 2005; Kondrak and Dorr, 2006; Kondrak and Sherif, 2006). He extended the work done by Covington to include multivalued phonetic features in his ALINE system (Kondrak, 2000). The task of phonetic alignment is difficult to evaluate as it requires expert knowledge of linguistics and the history of the languages in question. However, it can be evaluated indirectly by applying it to the task of cognate identification.

In the ALINE system, Kondrak used a series of multivalued phonological features that looked at the position in the mouth where the sounds were formed. Place of articulation can be easily converted into a multivalued scale that intuitively models similarity. Bilabial consonants are closer to labiodental consonants than they are palatal, for example. The same is true of vowels, which are measured in terms of height in the mouth and position from back to front. This similarity breaks down slightly when it comes to manner of articulation. He uses similarities for stops, affricates, fricatives and approximants in a way that is relatively intuitive for consonants, but also throws in values for vowels into the same feature. In addition to features, he uses salience based on a number of factors to weight the features. The salience values he uses are based on his intuitions and acknowledges that a principled manner for deriving them was an open question.

As in Covington's algorithm, Kondrak's produces many hypothesis alignments and must score each one. Whereas Covington opted for a more brute force approach, Kondrak relies on dynamic programming. He points out that while Covington saw the performance gain as negligible since most sequences were short, for a system to be applicable on a large scale it must be efficient. Kondrak evaluates his system on Covington's dataset of 82 cognate pairs in various languages. Comparing against the values Covington reported, Kondrak shows that his system is able to correct many of the mistakes and is clearly the better aligner. The extension of this work to cognate identification is given in Section 3.3.1. It is worth noting that this method is not empirical, but is driven by linguistic knowledge.

$P(w)$	Probability of English words
$P(e\|w)$	Probability of pronunciations for English words
$P(j\|e)$	Probability of Japanese sound for English sounds
$P(k\|j)$	Probability of *katakana* character for Japanese sounds
$P(o\|k)$	Probability of OCR errors for *katakana* characters

Figure 2: Probabilistic models for machine transliterating English and Japanese *katakana* (Knight and Graehl, 1998).

3.2.2 Machine Transliteration

Closely related to the issue of phonetic similarity is transliteration. Transliteration is the task of converting a word from the alphabet of one language to another with phonetic equivalents. Rather than being strict cognates, the words in the target language are borrowings (or proper nouns that do not make sense to translate, such as person names). Important work in this area was done by Knight and Graehl who developed a weighted finite state transducer (WFST) to convert Japanese *katakana* into English words (Knight and Graehl, 1998). Finite State Transducers have a long track record of successful application to tasks in computational morphology (Beesley and Karttunen, 2003). Also the task of transliteration as Knight and Graehl formulate it decomposes naturally into a cascade of finite state transducers, which have the handy property of being easily composable.

Knight and Graehl describe several challenges in producing their system. Transliterating into Japanese is relatively easy, but back-transliterating is more difficult. There is a one-to-many relationship between English words and *katakana* representations in Japanese. The problem is not as simple as just converting *katakana* characters to their Roman alphabet equivalents as there is a one-to-many mapping in that direction. To solve the problem they proposed a list of finite state transducers for each of the probabilistic models in Figure 2. The final model is specific to the task of scanned data, taking into account errors in optical character recognition (OCR). The first probabilistic model can be represented as a weighted finite state automaton since no transduction is necessary. The rest of the models were defined to be weighted finite state transducers. Here, the order is important. Each step in the figure must be fed downward as the transducer is composed.

Knight and Graehl used a corpus of short news articles from which they extracted 1449 unique *katakana* phrases, 222 of which were missing from an online bilingual dictionary. They back-transliterated these 222 phrases and found that most were either perfect or good enough. Another experiment looked at the names of 100 U.S. politicians in *katakana* that did not come from OCR. They tested the system using human subjects who were native English speakers and news aware. After being given brief instructions, they were set the task of back-transliterating these names. Their system got 76% correct or mostly

correct while the human subjects got only 34%.

Their system was not only successful, but it also has a useful implication for phonetic similarity. The probabilistic model $P(j|e)$ is essentially a measure of phonetic similarity. The probabilities issued by the WFST can be used as measures of phonetic similarity. The composability of the WFST also allows for methods that can work without having phonetic transcriptions available beforehand. Given sufficiently good transcription models, phonetic similarity (and therefore cognate identification) can potentially be calculated on much larger datasets.

3.3 Phonetic Methods

Phonetic methods to cognate identification seem to be well motivated by historical linguistic theory. The actual process of language change takes place in spoken form, which can be represented phonologically (though imperfectly). The earliest such work was done by Jacques Guy (Guy, 1984, 1994). In (Guy, 1984), Guy looked at semantically aligned bilingual word lists to find sound correspondences. To do this, he constructed a matrix of observed character frequencies for characters in the source and target languages across all word pairs. To produce a matrix of expected values for each correspondence, he considered the case where all correspondences are random. He then produced the matrix of expected values as follows:

$$E_{i,j} = \frac{\sum_{i=1}^{T} A_{row_i} \sum_{j=1}^{S} A_{column_j}}{\sum_{i=1}^{T} \sum_{j=1}^{S} A_{i,j}}, \tag{8}$$

where A is the matrix of observed frequencies, and T and S are the number of unique characters for each of the input languages. The character correspondences with the greatest differences in observed and expected frequencies are deemed correspondences. He deals with null correspondences by constructing matrices of weights and potentials for each word pair. Null correspondences occur when the length of words differ and thus the matrix is no longer square. Weights are the difference between observed and expected frequencies. Potentials are the sum of the weights and the maximum of all possible potentials to the right in the matrix. By ranking the characters in order of potential, null correspondences are values in the longer string that map to nothing in the shorter string.

Guy uses the notion of potentials as a measure of cognateness. The cognateness measure is the maximum potential of a character-to-character mapping in a word pair divided by the total number of correspondences (including null). He evaluates his technique using 300 words in 75 languages and dialects of Vanuatu. Languages that share more than 40% cognates and those with simple correspondences yield excellent results. Complex correspondences or low levels of cognate

sharing cause the results to deteriorate rapidly. Guy extends this work in (Guy, 1994) and attempts to construct a unified model of machine translation and diachronic phonology. This amounts to constructing a bilingual word list using cognates. The actual claim of creating a unified model is underdeveloped in the paper and is left as future work. Another criticism of this work is its heavy reliance on *ad hoc* parameters and thresholds. Guy's approach is probabilistic in that it uses co-occurrence statistics of actual data to determine correspondences. He then incorporates heuristics to fine-tune how these statistics are interpreted to produce assessments of cognateness.

3.3.1 Adding Semantic Similarity

The first attempt to derive cognates directly from unaligned vocabulary lists was done by (Kondrak, 2001). In this work, Kondrak combines phonetic similarity developed using the ALINE system (see Section 3.2.1) with a measure of semantic similarity. He describes the task as operating on two levels: word and language. On the word level he is computing a likelihood that a given word pair consists of cognates. On the language level he is computing which words are cognates given two phonetically transcribed vocabularies. The vocabularies must contain glosses to a metalanguage, such as English, in order to be able to compute semantic similarity. To do so, he uses WordNet to find all the synonyms, **hyponyms**, and **meronyms** for each gloss of a word. Hyponyms are words that are subclasses of larger classes. For example, *trout* is a hyponym of *fish*. Meronyms are words that are parts of something else. For example, *palm* is a meronym of *hand*. Regular string matching is done on the sets of possible cognate pairs to determine similarity, which is weighted based on the type of match. The phonetic similarity score generated by ALINE is interpolated with the semantic similarity score to produce the final likelihood score.

Kondrak evaluates his results on a dictionary of four Algonquian languages produced by Hewson (Hewson, 1993). The output of his system is a sorted list of suspected cognate pairs, where true cognates are more frequent near the top of the list. Determining where to threshold the list is application dependent, so Kondrak decided to calculate the average precision using three different threshold values (20%, 50%, and 80%). He calculates 3-point average precision for several different combinations of his methods as well as the simple string matching by Simard et al. (Simard et al., 1993), Dice's coefficient (Brew and McKelvie, 1996) and LCSR (Melamed, 1999). Kondrak found that his method with WordNet relations performed the best and significantly outperformed LCSR and the other orthographic methods. This result was very important as it showed that phonetic methods for cognate identification which incorporated linguistic knowledge could outperform string matching approaches.

3.3.2 Learning Approaches

A variety of machine learning approaches have recently been applied to the task of phonetic similarity and cognate identification. This section highlights

two approaches: Pair HMMs and Dynamic Bayesian Networks, both graphical models. Additional machine learning and statistical approaches include using support vector machines to predict orthography (Mulloni, 2007), applying Bayes theorem to cognate identification (Ellison, 2007), and semi-supervised learning of partial cognates (Frunza and Inkpen, 2006).

Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBNs) are graphical models which include Hidden Markov Models (HMMs). Filali and Bilmes describe a method using DBNs that both models context and learns edit distance to classify pronunciations (Filali and Bilmes, 2005). Their method develops a stochastic model of edit distance. Edit distance operations deal with the number of insertions, deletions and substitutions necessary to transform a source word into a target word. By assigning costs to each such operation, edit distance can be turned into a stochastic model with the probabilities learned from data. DBNs can then be used to represent this stochastic model and established learning techniques can be applied. Filali and Bilmes add memory to the model by changing the links in the DBN to condition the probability of the current operation on what the previous operation was. Additionally, they incorporate context-dependence by adjusting probabilities depending on what letters occur in the source or target words. They evaluated their system on its ability to learn edit distances and produce hypotheses for English pronunciations on the Switchboard corpus.

Pair HMMs

Pair Hidden Markov Models were introduced by MacKay and Kondrak to identifying cognates (Mackay and Kondrak, 2005). Previously they had been used in computational biology to align sequences of DNA (Durbin et al., 1998). The technique used by MacKay and Kondrak is to learn edit distance costs from data. The Pair HMM consists of a state for each edit operation. In a Pair HMM, there are two output streams. In the case of phonetic alignment, there is an output stream for each word that is being aligned. Their method manages to identify cognates with high precision and separates words that are similar due to chance. An advantage of this approach is that it is not domain-dependent and so may be applied to any situation requiring computation of word similarity. Wieling et al. applied this approach to Dutch dialect comparison (Wieling et al., 2007). Their results confirm MacKay and Kondrak's results and show that Pair HMMs align linguistic material well and produce reliable word distance measures. Fundamentally, this approach is equivalent to Filali and Bilmes (2005). The difference lies in the choice of representation (Pair HMMs versus the more general Dynamic Bayesian Networks) and the application of the method specifically to cognate identification.

Evaluation

Work by Kondrak and Sherif evaluate various techniques as a measure phonetic similarity applied to cognate identification (Kondrak and Sherif, 2006). They found that DBNs (Filali and Bilmes, 2005) outperformed all other approaches tested on average, which included Pair HMMs (Mackay and Kondrak, 2005), CORDI (Kondrak, 2002b), Levenshtein with learned weights (Mann and Yarowsky, 2001), ALINE (Kondrak, 2000), and several others. The dataset used was the Comparative Indo-European Data Corpus (Dyen et al., 1992a). The data consists of word lists of 200 meanings representing basic meanings of 95 languages and dialects belonging to the Indo-European language family. Kondrak and Sherif report extracting 180,000 cognate pairs from this corpus. Each system was evaluated based on cognate identification precision and averaged for the 11 pairs of the test set. The result of the evaluation was that graphical models (Pair HMMs and Dynamic Bayesian Networks) outperform manually designed systems when enough training data exists.

3.4 Future Work

Cognate identification is a central task in historical linguistics. It also has applications to dialectology and machine translation. It draws on measures of word similarity, techniques from computational phonology, and machine learning. We have described the two main methods used to identify cognates, each driven by different concerns. Orthographic methods were driven largely by the need to align bitexts for machine translation. Phonetic methods were originally driven by the need for cognate lists in historical linguistics. As time progressed and the application to machine translation became more obvious, the various techniques underpinning cognate identification have been driven by a greater variety of concerns. In recent years, application of machine learning approaches to cognate identification have become more popular.

Future work in cognate identification will most likely rely more heavily on statistical and machine learning approaches. Whereas in machine translation, the statistical MT paradigm has dominated the field, cognate identification has not yet reached this state. New models are being proposed and new methods for finding phonetic similarity, learning from real orthographies rather than phonetic transcriptions, and incorporating linguistic information are all open areas. One thing that has not yet been tried is an ensemble method that combines several different approaches to merge strengths of different approaches. Also, we expect cognate identification will be recognized as a useful tool in other areas of language technologies. One such area where it has been used and will probably see further exploration is in spelling correction. Kondrak and Dorr report favorable results applying ALINE and a combination of orthographic similarity measures to the problem of confusable drug names – instances where drug names look or sound alike and cause dangerous prescription errors (Kondrak and Dorr, 2006).

One major obstacle in the way of future progress is data sparsity. Human phonetic transcriptions are expensive to obtain and require considerable effort. Machine generated transcriptions are imperfect and contain noise that can interfere with the process of finding reliable sound correspondences. A leap forward in the accuracy of automatic phonetic transcriptions for a wide variety of orthographies would open the field of cognate identification to a large amount of data that it has previously been denied.

At first blush, the topic of cognate identification may seem mundane. Every learner of a second language is taught to look out for cognates at an early stage in their instruction. Strict cognates in the historical linguistic sense are much more than similar-looking or -sounding words. The task is one that requires information from a large variety of sources and has application to many different fields. As such, it will probably continue to grow as a field of active research for years to come.

4 Reconstructing the Evolutionary History

One of the primary aims of historical linguistics is to identify and establish historical relationships between languages. Cognate words between languages, recurrent sound changes that occured in a language's history and the similarities in the grammatical features provide evidence for these relationships (see Section 5.1.2). Traditionally a linguist looks at the evidence and tries to come up with a evolutionary tree that can explain as much of it as possible. All the major linguistic families have been established in this manner.

However the task of infering the underlying evolutionary structure becomes difficult as the number of languages under consideration increases since the linguist must keep track of more and more data. Often different linguists come up with similar but diffirent underlying structures and there is no objective way to evaluate them.

In this section and the next, we cover some of the work that seeks to formalize this process and develop methods for infering the underlying evolutionary structure based on the observed evidence.

4.1 Lexicostatistics

Lexicostatistics was introduced by Morris Swadesh in Swadesh (1950). Starting with a list of cognate words in the languages being analyzed, it builds a evolutionary tree for them. There are 3 main steps in applying Lexicostatistics.

Meaning List We start by choosing a universal and relatively culture-free core vocabulary list which we hope to be resistant to replacement or borrowing. Swadesh proposed a list of 200 such concepts including body parts, numerals, elements of nature, etc. – things which should be present in any language used for human communication. Once such a list has been chosen, it is filled with the most common words used for these concepts from all the languages that we want to analyze.

Finding Cognates Among the words corresponding to the same meaning slot, we identify cognate words by applying the comparative method.

Clustering the languages Lexicostatistics uses a distance based clustering method called UPGMA (Unweighted Pair Group Method with Arithmetic Mean). The distance between a pair of languages is measured by the percentage of shared cognates between them. The clustering algorithm proceeds like this:

1. Find the two closest languages (L1, L2) based on percentage of shared cognates.
2. Make L1,L2 siblings.
3. Remove one of them, say L1 from the set.
4. Recursively construct the tree on the remaining languages.
5. Make L1 the sibling of L2 in the final tree.

4.2 Glottochronology

Glottochronology is an application of Lexicostatistics which tries to estimate the time of divergence of siblings in the evolutionary tree. Glottochronology works under the assumption of the lexical clock.

Lexical Clock At all times the rate of lexical replacement is constant for all languages.

This constant is known as the glottochronological constant. Lee computed this constant to be .806 +/-0.0176 at 90% confidence level (Lees, 1953) based on 13 language pairs which are known to be related and for which times of divergence are known from historical records. This means that after 1000 years of divergence, two sibling languages will share approximately 81% of basic vocabulary.

Following Swadesh (1950), if t is the time in millennia, c is the percentage of cognates shared and r is the glottochronological constant, the time of divergence for any two languages can be computed using

$$t = \frac{\log c}{2 \log r}$$

.

4.3 Discussion

Lexicostatistics and Glottochronology have been criticized for their underlying assumptions of the lexical clock and for the difficulty of selecting a core vocabulary list.

Finding a list of 200 core concepts that would generalize across languages is a very hard task. Even the basic vocabulary of a language is affected by the culture. Some languages may fail to have any word for a particular meaning,

some may have multiple words for the same concept (synonyms, for example "small" and "little" in English) or the same word for more than one concept in the list (Ecuadorian Quechua has the same roots for "mouth" and "tongue") (McMahon and McMahon, 2006). Also, while resistant to borrowing, loan words can appear in the list. They must be detected and removed. Additionally, some kind of words are likely to be similar across languages irrespective of the evolutionary relationship (nursery words like mama, papa, imitations of sounds like bang, thud) (Kessler, 2001) and so must be removed. Taking these issues into consideration, Swadesh later proposed a 100 word list to be used for Lexicostatistics purposes (Swadesh, 1955b).

Other researchers have adapted the basic Swadesh list for different geographical regions like Southeast Asia (Matisoff, 2000), Australia (Alpher and Nash, 1999) leaving aside the lofty goal of building one list for all the languages of the world. These lists are more useful for the languages residing in one geographical area.

For Glottochronology, the main point of criticism has been the lexical clock assumption. It is well known that rates of lexical replacement are vastly different across languages (Old Armenian and Modern Armenian share 97% cognates (Bergsland and Vogt, 1962) while East Greenlandic Eskimo languages have system of taboo words which hastens the process of vocabulary loss (McMahon and McMahon, 2006)) and also for different words. Some work at adressing these issues is reported in Sankoff (1973); Brainerd (1970) and Embleton (1986).

The primary contribution of these techniques has been the idea of Swadesh lists. They provide a good baseline for compiling a list of words that can be reliably used for evolutionary analysis owing to their resistance to borrowing and their presence in most of the languages.

In last 5-10 years, however, many researchers have started looking at the phylogenetic methods developed in the evolutionary biology community and have applied that to language data. There has been work both in the lexicostatistics tradition where we only want to find the evolutionary tree and in the Gllottochronological tradition where we are interested in finding the dates of divergence events in tree. We survey these works in the next section.

5 Phylogenetic Inference

> In biology, phylogenetics (Greek: phyle = tribe, race and genetikos = relative to birth, from genesis = birth) is the study of evolutionary relatedness among various groups of organisms (e.g., species, populations). – Wikipedia : Phylogenetics

Evolutionary biology and historical linguistics address very similar problems. Inferring an evolutionary tree of the species/languages seen today and those that are extinct now, remains a key activity for both. In his book *The Descent of Man* , Darwin notes,

> "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."(pp. 89-90)

The processes of language evolution and species evolution are very similar. One ancestral language diverges to form two separate languages due to reasons like geographical distance, migration, mutation (linguistic innovations like sound change) and others. Ancestral languages become extinct with time and we come to know about them through any written record that they leave behind. Sanskrit and Latin are two such examples. The task in phylogenetic reconstruction is to look at the characteristics of present day and extinct languages and infer the underlying evolutionary tree.

The problem of phylogenetic reconstruction is well studied in evolutionary biology community. After the structure of DNA was identified in 1950s, the amount of data available to biologists became enormous. Over the last 40 years, many algorithmic and numerical methods have been developed that can handle such large amounts of data. Felsenstein (2003) provides a good starting point for an in-depth introduction to the general topic of phylogenetic inference and its usages.

In the following sections, we would describe some basic concepts and techniques for phylogenetic reconstruction and the attempts to apply them to language data.

5.1 Basic Concepts

Like any long standing scientific field of study, Phylogenetics has its own set of vocabulary. Following Ringe et al. (2002) and Warnow et al. (1996), we define the basic terminology in this section. Please note that *phylogenetic inference* and *phylogenetic reconstruction* refer to the same thing and may be used interchangeably in most cases.

5.1.1 Phylogenetic Trees and Networks

A phylogenetic tree represents a hypothesis about the possible evolution pattern of a set of languages. All the observed languages sit at the leaves of the tree. The internal nodes of the tree represent the ancestral languages that are not observed. The length of a tree branch may represent the time since the divergence. In the tree, the languages are assumed to evolve independently after the event of divergence.

These trees can be rooted or unrooted. An unrooted tree represents the relationships between the languages but does not identify the ancestry. A rooted tree also identifies the common ancestors of the related languages. Determination of the root of an unrooted tree often needs some extra information which may or may not be available.

languages	English	German	French	Italian	Spanish	Russian
words	hand	Hand	main	mano	mano	ruká
states	1	1	2	2	2	3

Table 1: Lexical character corresponding to Hand (Based on data from Ringe et al. (2002))

Phylogenetic networks are a generalization of trees that allow for contact between languages after they have diverged, represented by horizontal edges between branches. This allows for explicit representation of language contact and borrowings.

5.1.2 Characters

Evolution of a language can be seen as a change in some of its features. These features may be a lexical item, a grammatical feature or a change in some sound. A character encodes the similarity between different languages on the basis of these features and defines a equivalence relation on the set of languages L.

Character A character is a function $c : L \to Z$ where Z is the set of integers and L is the set of languages. (Warnow et al., 1996)

These characters can take one or more values which are called the states of the character. Every language can be represented as a vector of character states. The actual values of the charcater states are not important. (Ringe et al., 2002)

Three main types of characters are used in linguistics.

Lexical A Lexical character corresponds to a meaning slot. Cognate classes of the words associated with the meaning slot in various languages form the states of the character. So two languages have the same state for the character if their words for this meaning slot are cognates. A example taken from Ringe et al. (2002) is shown in Table 1. Here the character encodes the lexical character for *hand*. English and German words for *hand* are cognates as are the words in French, Italian and Spanish. Russian word is not cognate to any of them.

Morphological Morphological characters are similar to lexical characters but instead of words, they represent inflectional markers and are coded by cognation. For example, English and German will have different states for the character *future tense* since they use different auxilary verbs in that construction but Spanish, Italian and French have the same construction coming down from the Latin and so they all have the same state.

Phonological Phonological characters represent sound changes. Since the possibility of a sound change reversing over the course of linguistic evolution is small, they are coded only by recording if the sound change has happened for the language or not. So they can only have two states. An example

is Grimm's Law which describes a series of sound changes in Germanic languages. So a voiced stop (bh) changed to voiceless stop (b) in English (brother), Dutch (broeder) but not in Sanskrit (bhrata). So English and Dutch will show the same state for this character and Sanskrit, the other.

Choosing a set of characters and properly coding them is an important step but unfortunately it still remains a black art. Researchers differ on how a particular character should be encoded (see Gray and Atkinson (2003); Evans et al. (2004)) and final trees obtained from various methods have been shown to be sensitive to the choice of characters Nakhleh et al. (2005b).

5.1.3 Homoplasy

Two languages can exhibit the same state for a character for reasons other than an evolutionary relationship. The same state may occur due to independent parallel development or due to back mutation. These cases are called homoplasy.

Parallel Development Two languages can independently evolve in the same manner. For example, the sound changes that simplify the consonant clusters are likely to happen in many languages on their own even when the languages are not related. If such characters are not detected, they may suggest a connection when there is none.

Back Mutation This refers to the phenomenon when a character evolves to a state which was already seen before in the tree i.e a old state of the character reappears. This is quite rare for linguistic characters described above.

In the absence of homoplasy, every character provides unambiguous information about the splitting order of languages and hence building the underlying tree is relatively easier. Most of the early work on linguistic phylogenetic inference assumes a homoplasy free evolution (Warnow et al., 1996; Ringe et al., 2002; McMahon and McMahon, 2006). It should be noted that borrowing between languages is not considered homoplasy and is another source of ambiguity. However, loan words can be accurately identified by the comparative method. Most studies throw away the characters that are suspect of borrowing (Ringe et al., 2002; Gray and Atkinson, 2003). Warnow et al. (2004b); Nakhleh et al. (2005a) have proposed models that take into account homoplasy and borrowing explicitly.

5.1.4 Perfect Phylogenies

When a character evolves down the tree without homoplasy (i.e. without parallel development and back mutation), it is said to be compatible on the tree. This implies that every time the state of a character changes, it enters a new state previously unseen on the tree up to that point in time. As a result, all the languages exhibiting the same state for a character form a connected subtree.

A tree on which all the characters under consideration are compatible is called a perfect phylogeny. A survey of the perfect phylogeny problem and approaches towards it is presented in Fernï¿½dez-Baca (2001). Also see Warnow et al. (2004a).

5.1.5 Models of Evolution

A model of evolution describes the process under which characters evolve on the tree. Not all methods of phylogenetic reconstruction need a explicit model of evolution. However parametric statistical methods like Maximum Likelihood and Bayesian Methods need them. In historical linguistics, both the varieties have been used. While Warnow et al. (1996); Gray and Jordan. (2000); Dunn et al. (2005) have used the earlier variety of models, Gray and Atkinson (2003); Pagel and Meade (2006) used a restriction site model of evolution that explicitly models rate of change of characters within a Bayesian framework.

Warnow et al. (2004a,b) discuss what kinds of models of evolution are appropriate for use with linguistic data and what kind of inference they allow. In particular, it is relatively easy to reconstruct unrooted tree under most models of evolution but only highly restrictive models allow for the estimation of divergence times and dates on tree nodes. Evans et al. (2004) also discusses the problem of inference of divergence times.

5.2 Methods for Phylogenetic Reconstruction

Once a set of characters has been chosen and a model of evolution decided upon, there are many different methods that can be used for reconstructing phylogenies. For a given set of languages and characters, all the possible evolutionary trees form a "tree space" and a phylogenetic inference method can be seen as searching for the optimum tree according to its optimization criteria.

These methods fall into two broad categories, distance based methods and character based methods.

5.2.1 Distance Based Methods

Distance based methods work by measuring some kind of distance between languages as described by the character states and putting the languages close to each other in the same subtree.

UPGMA The simplest distance based method is UPGMA (Unweighted Pair Group Method with Arithmetic mean). It recovers the correct tree if the input data satisfies the lexical clock hypothesis (see Section 4.2). The method was described in Section 4.1.

Neighbor-Joining NJ (Saitou and Nei, 1987) is a greedy algorithm like UPGMA but it doesn't need the lexical clock assumption to retrieve the correct tree. A distance matrix between all pairs of languages is given as input to the algorithm. The method starts out with a star-like topology

and at every step tries to minimize an estimate of total length of the tree by combining together the languages that provide the most reduction. It has been shown that the method is statistically consistent (i.e. if there is a tree on which the input distances fit perfectly, it will recover that tree). The exact criterian that NJ tries to optimize has also been established. See Gascuel et al. (2006) for further details.

5.2.2 Character Based Methods

Character based methods work with states of characters rather than just looking at the total number of changes of character states between languages.

Maximum Parsimony Maximum Parsimony (MP) is based on the principle of minimum evolution. It seeks to find trees on which minimum number of character state changes occur. There are many different measures of parsimony with Fitch parsimony (Fitch, 1971) and Wagner parsimony (Farris, 1970) being two frequently used examples. Finding a MP tree is NP-Hard and so heuristic searches are used. MP usually returns more than one tree – none of which, however is guaranteed to be optimal.

MP is a non parametric statistical method since it does not induce a model of character evolution along the branches. However it is not statistically consistent (Felsenstein, 1978). It suffers from what is known as *long branch attraction*. This means that if the rates of evolution are very different on different branches of the true tree then MP is likely to reconstruct the wrong tree.

For use of MP in constructing language phylogenies, see Gray and Jordan. (2000); Dunn et al. (2005); Ryder (2006).

Maximum Compatibility Maximum Compatibility (MC) seeks to find the tree on which most number of characters are compatible (5.1.4). If we assume that evolution is homoplasy free, then all the characters in the input data should be compatible on the true tree. Hence the problem reduces to finding a perfect phylogeny which is known to be NP-Hard (Bodlaender et al., 2000). However if the maximum number of states per character is bounded then it is possible to find the best tree in polynomial time (Kannan and Warnow, 1997).

In practice however it is not always possible to find a perfect phylogeny on the given character data. In that case, MC returns the tree(s) that has maximum number of characters compatible on it. Like MP, MC also does not induce a model of character evolution along edges.

Ringe et al. (2002) use MC to construct the phylogenetic tree for IE data set.

Maximum Likelihood The Maximum likelihood (ML) method seeks to find the tree that maximizes the likelihood of the observed data under an

assumed parametric model of character evolution. ML is statistically consistent and is resistant to noise in the data and differing rates of evolution along different branches. However it is only as good as the underlying model of evolution. Moreover, since it involves search over two spaces, the parameter space of the model of character evolution and the space of tree topologies, it is very expensive in practice.

J. (1981) describes a dynamic programming algorithm for compuitng $P(D|\theta, T)$ i.e. the probability of observed data given parameters of the model and the tree topology. A variation of the structural EM algorithm for estimating θ and T simultaneously is described in Friedman et al. (2002).

Bayesian Inference This is a more recent addition to the stable of phylogenetic reconstruction methods. The aim here is to estimate the posterior probability distribution of trees given a prior and the observed character data. Like ML, this needs a model of character evolution. Since it is hard to compute the posterior distribution explicitly, Markov Chain Monte Carlo (MCMC) is used to approximate the posterior distribution of trees. For more details, see Huelsenbeck et al. (2002).

The most attractive part of the Bayesian framework from the linguistic point of view is the possibility of including a prior. Linguistic data is often very small in amount and so parameter estimation for richer models of evolution using ML is hard. In such cases, BI is more suitable. Moreover, priors allow us to bring the evidence available from other fields like genetics, history, sociolinguistics into the picture that provide important information about the evolution process.

BI has been used by Gray and Atkinson (2003); Pagel et al. (2007); Atkinson et al. (2005) to construct a evolutionary tree for IE data set.

5.2.3 Comparison of various methods

Given all the methods described in the last section, the obvious question is which method works best for language data or more specifically, which method works best under what conditions? The first question can be answered by testing the methods on some data set for which we know the true evolutionary tree with some certainity. To answer the second question, we need controlled experiemnts where effect of one factor could be teased apart from the other. Nakhleh et al. (2005b) and Barbancon et al. (2007) present results of comparing many of the methods described above. While Nakhleh et al. (2005b) compare them on Indo-European data set which is a well studied language family and the true evolutionary tree is known to a large extent, Barbancon et al. (2007) did experiemnts with syntheic data. In these experiemnts they could control for various factors like contact between languages, degree of homoplasy, difference in rate of change of characters, choice of characters and see their effect on different methods.

	UPGMA	BI	WMC	MC	MP	NJ
Full(336)	115	51	53	48	52	53
Lex(297)	98	43	44	44	45	44
Full-screened(294)	75	15	15	14	14	17
Lex-screned(259)	61	12	9	9	9	10

Table 2: No of characters incompatible on trees returned by different methods on IE data set (Data taken from Nakhleh et al. (2005b))

Nakhleh et al. (2005b) They used the data collected by Ringe et al. (2002). It consisted of 336 characters for 24 IE languages. They did experiemnts under 4 data conditions namely all (all the characters), lex(only lexical characters), all-screened , lex-screened (characters exhibiting parallel development are removed). They evaluated the reults based on how many characters were incompatible (see Section 5.1.4) on the final tree returned by the method. Results are shown in Table 2. In general, all the methods perform better when data is screened and UPGMA does the worst. For a further discussion of the results, see Nakhleh et al. (2005b).

Barbancon et al. (2007) They first generated a random binary tree under a generative model. Then using the evolution model described in Warnow et al. (2004b), characters were evolved along the tree branches. Homoplasy and deviation from lexical clock were explicitly controlled. The error rate was measured by false negatives which is the count of bi-partitions of leaves that exist in the original tree but not in the tree proposed by the method and false positives which is the count of bi-partitions that occur in the tree proposed by the method but not in the true tree. They tested the same set of methods as Nakhleh et al. (2005b) under various settings of parameters. Their main conclusion was that some methods like MP and MC are able to take advantage of data that is free from homoplasy while others like BI perform more or less the same. for further results, please see Barbancon et al. (2007).

5.3 Modeling Homoplasy and Borrowing

As mentioned in Section 5.1.1, phylogenetic networks are a generalization of tree typology that allow for explicit modelling of language contact and borrowing events. It is possible to adapt the methods described in the last section for network reconstruction (see Jin et al. (2006b,a); Nakhleh et al. (2005a); Bryant and Moulton (2002); Bandelt et al. (1999)).

Extending the work of Ringe et al. (2002), Nakhleh et al. (2005a) presented perfect phylogeny networks as a generalization of perfect phylogeny problem on trees. We said in section 5.2.2 that it is not possible to always find a perfect phylogeny on a tree. The idea here is to add the minimum number of horizontal contact edges to the tree obtained by applying MC that will make all of the

characters compatible on the tree. The key assumption in their model is that when a word is borrowed into a language, it replaces the original word completely (i.e. borrowing across contact edges does not lead to polymorphic characters) [5]. Working with the same data used by Ringe et al. (2002), they were able to obtain a perfect phylogeny network by adding 3 additional contact edges to the tree obtained by Ringe et al. (2002). These contact edges were consistent with geographical and chronological constraints on possible contacts.

There has also been work on developing models of character evolution that explicitly account for homoplasy. Such models have been porposed by Warnow et al. (2004b); Atkinson et al. (2005).

5.4 Dating of nodes in Phylogenetic Trees

As we have mentioned above, inferring tree topology is possible under most models of evolution. However to estimate the times associated with internal nodes, we need additional assumptions on rates of evolution of characters.

Extending the work in Gray and Atkinson (2003) , Atkinson et al. (2005) present results on estimating the dates of divergence on both real IE data and on synthetic data.

Their basic model is a two state time-reversible model of lexical evolution. Now there is variation in the rates of evolution across different lexical items. This is modelled with a gamma distribution over the rates of evolution. The rates can also vary through time. To handle this, they use the penalized-likelihood rate smoothing that penalizes any abrupt changes in the evolution rates on adjacent branches. Using BI, they generated a sample distribution of trees from the posterior probability distribution under this model.

To estimate the absolute time of divergence, they constrained the range of dates on 14 of the nodes in the tree based on available historical evidence. The estimate of dates that they obtained supports Anatolian theory of Indo-European origin (Atkinson and Gray, 2006). These estimates were found to be robust towards choice of character data and priors. The authors used BI to also estimate dates on the same data set but under a very different model of evolution. The date estimates obtained were consistent with those obtained earlier.

However other authors have consistently argued against inference of dates on phylogenetic trees. In particular Evans et al. (2004) have proposed a *no common mechanism model* in which all the rates of evolution are assumed to be independent and no inference of times of divergence is possible. They consider the attempts to infer the divergence dates as premature and advise against it.

5.5 Looking for deeper relationships

Methods of phylogenetic reconstruction that use lexical characters require that the cognate words be identified between the languages being analyzed. However

[5] a character for which some language exhibits more than one state

the method of cognate identification gives reliable results only upto a time depth of $8,000 \pm 2,000$ years (Nichols, 1992). After that point in time, it becomes hard to distinguish chance resemblance from resemblance due to shared lineage. However, some recent work has shown that it may be possible to construct phylogenies beyond a time depth of 10,000 years.

Some researchers (Dunn et al. (2005); Ryder (2006)) have experimented with structural (grammatical) features of languages as the the basis for phylogenetic reconstruction. These features are encoded as binary characters based on the presence and absence of the feature in a language.

Dunn et al. (2005) used structural features to try and reconstruct a phylogeny of Papuan languages of Island Melanesia. These languages exist in close contact with around 100 languages belonging to Austronesian family. Dunn et. al. found that these languages share almost no lexical cognates other than the borrowings from the neighboring Austronesian languages. Based on the rate of lexical replacement seen in other languages, this suggested that either these languages were completely unrelated or they diverged a long time ago. However they did share many structural properties. So the authors collected data about 125 structural features from 15 Papuan and 16 Austronesian languages, taking care to avoid features that are known to occur together in languages with high correlation.

To confirm their hypothesis,they first applied the Maximum Parsimony method to Austronesian languages and obtained a tree congruent to the one prepared by the comparative method. This showed that grammatical characters can also reliably extract phylogenetic relations. Then they applied the same technique to the Papuan languages. As expected, the strength of phylogenetic signal was relatively weak in this case but it produced a geographically consistent tree showing broad genological groupings of the languages.

In Ryder (2006), the author has used the data available from the World Atlas of Languages (). He found that Bayesian Inference performed better than Maximum Parsimony for structural features.

Another effort at pushing back the time depth accessible to phylogenetic reconstruction methods is in Pagel et al. (2007) where they identify a consistent connection between frequency of word usage and rate of word evolution. Using the bayesian framework of Gray and Atkinson (2003), they inferred a phylogenetic tree of 87 Indo-European languages. Similar to Pagel and Meade (2006), they estimated lexical evolution rates for a list of 200 basic meanings as the mean of the Bayesian posterior probability distribution of evolution rates. These rates were analyzed against the frequency of use of these words. These two quantities showed a negative correlation across 4 languages they tested on. This was consistent across part of speech and accounted for almost 50% variation in the rates of evolution. This shows that the words that are used most are least susceptible to change.

In the light of these developments, Russell (Gray, 2005) suggests that it may be possible to detect a phylogenetic signal even for time depths greater than 10,000 years. This opens up doors for possible testing of long distance language relationships proposed in linguistics literature.

5.6 Working with raw language data

Most of the work on phylogenetic inference in linguistics has been done based on lexical characters that are extracted by human linguists by applying the camparative method. Indo-European datasets prepared by Dyen et al. (1992b) and Ringe et al. (2002) have been used in many of these studies. However obtaining such data for new language families remains a time consuming and hard task. This limits the utility of computational phylogenetic methods since they can only be used on well analyzed sets of languages.

Multilateral comparison (see section 2) on the other hand works directly with word lists in various languages. That makes it interesting from a computational point of view since the more data we give as input, the more likely a consistent statistical method would be to reconstruct the correct tree.

Kessler (2001) suggests that the primary difference between the comparative method and multilateral comparison is of statistical significance. When do we have enough evidence to proclaim relatedness? Untill now it has been a matter of the gut feeling of the linguists but using strict statistical methodology can bring in objectivity.

In section 3, we discussed many ways of automaticaly identifying the cognates however they have never been used in a phylogenetic reconstruction study untill now. Recently, Bouchard et al. (2007) have used the automatically extracted cognates to learn the sound changes that may have happened during the evolution from the ancestral language. In their model, words undergo stochastic edits along the branches of the underlying tree. These edits are context sensitive. They fix the topology of the underlying phylogenetic tree and use EM to estimate the parameters of the stochastic process (see the paper for more details). They use this setup for a number of tasks like identifying phonological changes and reconstructing word forms at the internal nodes of the underlying tree.

This looks like a promising direction since it uses large corpora but is not limited to surface matching. By learning the regular sound correspondences, it incorporates the strong point of comparative method in the computational framework. It can also act as a cognate identification algorithm similar in spirit to methods that look at the phonetic similarity between words to identify cognates (see section 3.2)

5.7 Future Directions

Quantitative methods for constructing linguistic phylogeny have just arrived on the scene. Their success untill now has mostly been limited to reproducing the already known phylogenies. Before they can be trusted with identifying true phylogeines, they need to be better understood in terms of their applicability under various scenarios of language change. Recent studies comparing various methods under different conditions provide a good starting point in that direction.

Pagel et al. (2007) has given an important insight into the mechanism of

lexical evolution which can be used to improve the models of evolutions used in Maximum Likelihood and Bayesian Inference. Understanding the evolutionary process of other types of language characteristics like syntax, word order, phonology, etc. remains an open area of exploration.

Another promising direction seems to be a computational framework for replacing the human involvement in cognateness judgments and working with larger corpora directly. A joint inference of cognateness and the sound changes under the framework of Bouchard et al. (2007) is a possibility, something that authors already seem to be looking at. As they have noted, their edit model is quite basic and does not capture many interesting phenomenan that happen in the real world. Developing richer models would be a natural next step.

6 Conclusion

We have presented a picture of two major areas where computational methods have been applied to historical linguistics. We began with cognate identification, a central task in historical linguistics with applications in machine translation and dialectology. Cognate identification uses techniques involving word similarity, computational phonology, and machine learning. The two main branches of cognate identification research were driven by their separate origins. Orthographic methods largely emerged from applications dealing with aligning bitexts. Phonetic methods originated mainly from a focus on historical linguistics. Recently these two approaches have begun to merge and machine learning techniques are playing a larger role. The end result is that cognate identification is a richer topic for research with many possible extensions and much work still to be done.

The other area seeing more and more computational techniques being used is inferring phylogenies of languages. This is a task that has a lot of similarities to the task of biological phylogenetic inference and faces much of the same problems. Recent years have seen more adaptions of phylogenetic inference techniques to linguistic data with varying degrees of success. However, most of the work has been a reaffirmation of what linguists have worked out through the comparative method. Also, we have yet to see techniques and evolutionary models that are specifically suited to linguistic data. Recent research uncovering relationships between word usage frequency and lexical evolution rates look like steps in the right direction. An interesting development has been efforts to work with raw language data instead of hand collected cognateness judgments. We feel that this line of investigation holds much promise for practitioners of computational techniques and will see more work in the future.

7 Acknowledgment

We would like to thank Mark Pagel for making his papers available to us.

References

Barry Alpher and David Nash. Lexical replacement and cognate equilibrium in australia. *Australian Journal of Linguistics*, 19 (1):5–56, 1999.

Q. D. Atkinson and R. D. Gray. How old is the indo-european language family? illumination or more moths to the flame? In P. Forster J. Clackson and C. Renfrew, editors, *Phylogenetic methods and the prehistory of languages*, chapter 8, pages 91–109. MacDonald Institute, Cambridge, 2006.

Quentin Atkinson, Geoff Nicholls, David Welch, and Russell Gray. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society*, 103(2):193–219, August 2005. URL `http://dx.doi.org/10.1111/j.1467-968X.2005.00151.x`.

H. J. Bandelt, P. Forster, and A. Rohl. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology Evolution*, 16(1):37–48, 1999.

Francois Barbancon, Tandy Warnow, Steven N. Evans, Donald Ringe, and Luay Nakhleh. An experimental study comparing linguistic phylogenetic reconstruction methods. Technical Report 732, Department of Statistics, University of California, Berkeley, May 2007.

K. R. Beesley and L. Karttunen. *Finite State Morphology*. CSLI Publications, Stanford, CA, 2003.

K. Bergsland and H. Vogt. On the validity of glottochronology. *Current Anthropology*, 3:115–153, 1962.

Hans L. Bodlaender, Michael R. Fellows, Michael T. Hallett, H. T. Wareham, and Tandy J. Warnow. The hardness of problems on thin colored graphs. journal version: The hardness of perfect phylogeny, feasible register assignment and other problems on thin colored graphs. *Theoretical Computer Science*, 244:167–188, 2000.

Alexandre Bouchard, Percy Liang, Thomas Griffiths, and Dan Klein. A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 887–896, 2007. URL `http://www.aclweb.org/anthology/D/D07/D07-1093`.

Barron Brainerd. A stochastic process related to language change. *Journal of Applied Probability*, 7 (1):69–78, 1970.

C. Brew and D. McKelvie. Word-pair extraction for lexicography. *Proceedings of NeMLaP*, 96:45–55, 1996.

David Bryant and Vincent Moulton. Neighbornet: An agglomerative method for the construction of planar phylogenetic networks. In *WABI '02: Proceedings of the Second International Workshop on Algorithms in Bioinformatics*, pages 375–391, London, UK, 2002. Springer-Verlag.

K.W. Church. Char_align: a program for aligning parallel texts at the character level. *Proceedings of the 31st conference on Association for Computational Linguistics*, pages 1–8, 1993.

M.A. Covington. An algorithm to align words for historical comparison. *Computational Linguistics*, 22(4):481–496, 1996.

M. Cysouw and H. Jung. Cognate Identification and Alignment Using Practical Orthographies. *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 109–116, 2007.

Michael Dunn, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075, September 2005. doi: 10.1126/science.1114615. URL `http://www.isrl.uiuc.edu/ amag/langev/paper/dunn05ancientLanguageSCIENCE.html`.

Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998.

I. Dyen, J.B. Kruskal, and P. Black. *An Indoeuropean Classification: A Lexicostatistical Experiment*. American Philosophical Society, 1992a.

Isadore Dyen, Joseph Kruskal, and Paul Black. An indoeuropean classification, a lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82, 1992b.

Mark T. Ellison. Bayesian Identification of Cognates and Correspondences. *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, 2007.

Sheila Embleton. Statistics in historical linguistics. *Quantitative Linguistics*, 30:(monograph), 1986.

S.N. Evans, Don Ringe, and Tandy Warnow. Inference of divergence times as a statistical inverse problem. In *Phylogenetic Methods and the Prehistory of Languages*. Cambridge, UK, July 2004. URL `http://www.isrl.uiuc.edu/ amag/langev/paper/evans04PhylogeneticMethods.html`.

J. Farris. Methods for computing wagner trees. *Systematic Zoology*, 34:21–24, 1970.

J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–410, 1978.

Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, September 2003. ISBN 0878931775. URL `http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0878931775`.

D. Fernï¿½dez-Baca. The prefect phylogeny problem. In X. Cheng and D.-Z. Du, editors, *Steiner Trees in Industry*, pages 203–234. Kluwer, 2001.

K. Filali and J. Bilmes. A Dynamic Bayesian Framework to Model Context and Memory in Edit Distance Learning: An Application to Pronunciation Classification. *Ann Arbor*, 100, 2005.

W. M. Fitch. Towards defining the course of evolution: minimal change for a specific tree topology. *Systematic Zoology*, 20:406–416, 1971.

N. Friedman, M. Ninio, I. Pe'er, and T. Pupko. A structural em algorithm for phylogenetic inference. *Journal of Computational Biology*, 9:331–353, 2002.

O. Frunza and D. Inkpen. Semi-supervised learning of partial cognates using bilingual bootstrapping. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 441–448, 2006.

Gascuel, Olivier, Steel, and Mike. Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11):1997–2000, November 2006. URL `http://dx.doi.org/10.1093/molbev/msl072`.

R. D. Gray and F. M. Jordan. Language trees support the express-train sequence of austronesian expansion. *Nature, 405*, pages 1052–1055, 2000.

Russell D. Gray. Pushing the time barrier in the quest for language roots. *Science*, 309(5743):2007–2008, September 2005. doi: 10.1126/science.1119276. URL `http://www.isrl.uiuc.edu/ amag/langev/paper/gray05languageRootsSCIENCE.html`.

Russell D. Gray and Quentin D. Atkinson. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426(6965):435–439, November 2003. doi: 10.1038/nature02029. URL `http://www.isrl.uiuc.edu/ amag/langev/paper/gray03languageTreeDivergence.html`.

J. Guy. An algorithm for identifying cognates between related languages. *COLING-84*, pages 448–451, 1984. URL `"citeseer.ist.psu.edu/guy84algorithm.html"`.

J.B.M. Guy. An Algorithm for Identifying Cognates in Bilingual Wordlists and its Applicability to Machine Translation. *Journal of Quantitative Linguistics*, 1(1):35–42, 1994.

J. Hewson. *A Computer-generated Dictionary of Proto-Algonquian*. Canadian Museum of Civilization, 1993.

Henry M. Hoenigswald. *Language Change and Linguistic Reconstruction*. University of Chicago Press, 1960.

J. P. H. Huelsenbeck, B. Larget, R. E. Miller, and F. Ronquist. Potential applications and pitfalls of bayesian inference of phylogeny. *Systematic Biology*, 51:673–688, 2002.

D. Inkpen, O. Frunza, and G. Kondrak. Automatic Identification of Cognates and False Friends in French and English. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 251–257, 2005.

Felsenstein J. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Maximum likelihood of phylogenetic networks. *Bioinformatics*, 22:2604–2611, 2006a.

G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Efficient parsimony-based methods for phylogenetic network reconstruction. *Bioinformatics*, 23:e123–e128, 2006b.

Sampath Kannan and Tandy Warnow. A fast algorithm for the computation and enumeration of perfect phylogenies. *"SIAM Journal of Computing"*, 26 (6):1749–1763, December 1997.

B. Kessler. *The Significance of Word Lists*. CSLI Publications, Stanford, CA, 2001.

K. Knight and J. Graehl. Machine Transliteration. *Computational Linguistics*, 24(4):599–612, 1998.

G. Köbler. *Germanisches Wörterbuch*. Arbeiten zur Rechts-und Sprachwissenschaft Verlag, 1980.

G. Kondrak. A new algorithm for the alignment of phonetic sequences. *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 288–295, 2000.

G. Kondrak. Identifying cognates by phonetic and semantic similarity. *North American Chapter Of The Association For Computational Linguistics*, pages 1–8, 2001.

G. Kondrak. Phonetic Alignment and Similarity. *Computers and the Humanities*, 37(3):273–291, 2003a.

G. Kondrak. Identifying Complex Sound Correspondences in Bilingual Wordlists. *Proceedings of CICLING 2003*, 2003b.

G. Kondrak. Combining Evidence in Cognate Identification. *Proceedings of Canadian AI 2004*, pages 44–59, 2004.

G. Kondrak. Cognates and Word Alignment in Bitexts. *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 305–312, 2005.

G. Kondrak. *Algorithms for Language Reconstruction*. PhD thesis, University of Toronto, 2002a.

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

G. Kondrak and T. Sherif. Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification. *Proceedings of the Workshop on Linguistic Distances*, pages 43–50, 2006.

G. Kondrak, D. Marcu, and K. Knight. Cognates can improve statistical translation models. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 46–48, 2003.

Grzegorz Kondrak. Determining recurrent sound correspondences by inducing translation models. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002b. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1072228.1072244.

Robert. Lees. The basis of glottochronology. *Language*, 29 (2):113–127, 1953.

W. Mackay and G. Kondrak. Computing Word Similarity and Identifying Cognates with Pair Hidden Markov Models. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 40–47, 2005.

G.S. Mann and D. Yarowsky. Multipath translation lexicon induction via bridge languages. *North American Chapter Of The Association For Computational Linguistics*, pages 1–8, 2001.

J. A. Matisoff. On the uselessness of glottochronology for the subgrouping of tibeto-burman. *Time Depth in Historical Linguistics*, pages 333–71, 2000.

T. McEnery and M. Oakes. Sentence and word alignment in the CRATER Project. *Using Corpora for Language Research*, pages 211–231, 1996.

April McMahon and Robert McMahon. *Language Classification by Numbers*. Oxford University Press, Oxford, 2006.

I.D. Melamed. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130, 1999.

A. Mulloni. Automatic Prediction of Cognate Orthography Using Support Vector Machines. *Proceedings of the ACL 2007 Student Research Workshop*, pages 25–30, 2007.

A. Mulloni and V. Pekar. Automatic Detection of Orthographic Cues for Cognate Recognition. *Proceedings of LREC'06, 2387*, 2390, 2006.

L. Nakhleh, D. Ringe, and T. Warnow. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *LANGUAGE, Journal of the Linguistic Society of America, 81(2).*, pages 382–420, 2005a.

L. Nakhleh, T. Warnow, D. Ringe, and S.N. Evans. A comparison of phylogenetic reconstruction methods on an ie dataset. *Transactions of the Philological Socity*, 2005b.

Johanna Nichols. *Linguistic Diversity in Space and Time*. The University of Chicago Press, Chicago and London, 1992.

M Pagel and A Meade. Estimating rates of lexical replacement on phylogenetic trees of languages. *Phylogenetic methods and the prehistory of languages (Peter Forster and Colin Renfrew eds.). McDonald institute Monographs*, pages 173–182, 2006.

Mark Pagel, Quentin D. Atkinson, and Andrew Meade. Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, 449(7163):717–720, Oct 2007. doi: 10.1038/nature06176. URL `http://www.isrl.uiuc.edu/ amag/langev/paper/pagel07wordFrequencyNATURE.html`.

D. Ringe, Tandy Warnow, and A. Taylor. Indo-european and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129, 2002. doi: 10.1111/1467-968X.00091. URL `http://www.isrl.uiuc.edu/ amag/langev/paper/ringe02IECladistics.html`.

D.A. Ringe. *On Calculating the Factor of Chance in Language Comparison*. American Philosophical Society, 1992.

Robin J Ryder. Grammar and phylogenies. Technical report, Department of Statistics, University of Oxford, 2006.

N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4:406–425, July 1987. URL `http://mbe.oxfordjournals.org/cgi/content/abstract/4/4/406`.

D. Sankoff. Mathematical developments in lexicostatistic theory. *Current Trends in Linguistics XI*, pages 93–113, 1973.

M. Simard, G.F. Foster, and P. Isabelle. Using cognates to align sentences in bilingual corpora. *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*, pages 1071–1082, 1993.

Noah A. Smith. Detection of Translational Equivalence. Technical report, University of Maryland, 15 May 2001.

M. Swadesh. Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, 21(2):121–137, 1955a.

Morris Swadesh. Salish internal relationships. *International Joural of American Linguistics*, 16:157–167, 1950.

Morris Swadesh. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21:121–137, 1955b.

J. Tiedemann. Automatic construction of weighted string similarity measures. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

T. Warnow, D. Ringe, and A. Taylor. Reconstructing the evolutionary history of natural languages. In *SODA: ACM-SIAM Symposium on Discrete Algorithms (A Conference on Theoretical and Experimental Analysis of Discrete Algorithms)*, pages 314–322, 1996. URL `citeseer.ist.psu.edu/article/warnow96reconstructing.html`.

T. Warnow, S.N. Evans, D. Ringe, and L. Nakhleh. Stochastic models of language evolution and an application to the indo-european family of languages. Technical report, Department of Statistics, The University of California, Berkeley, 2004a. URL `http://www.isrl.uiuc.edu/ amag/langev/paper/warnow04stochasticModels.html`.

T. Warnow, S.N. Evans, D. Ringe, and L. Nakhleh. A stochastic model of language evolution that incorporates homoplasy and borrowing. In *Phylogenetic Methods and the Prehistory of Languages*. Cambridge, UK, July 2004b. URL `http://www.isrl.uiuc.edu/ amag/langev/paper/warnow04PhylogeneticMethods.html`.

M. Wieling, T. Leinonen, and J. Nerbonne. Inducing Sound Segment Differences Using Pair Hidden Markov Models. *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 48–56, 2007.