19,050 words

Automatic language similarity comparison using n-gram analysis.

Ben Coppin
Queens' College
June 2008

This dissertation is submitted for the degree of Master of Philosophy

CONTENTS

1 Preface

1.1 Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

1.2 Structure of this thesis

- Section 3: Literature review; the history of lexicostatistics.
- Section 4: Review of the reliance of lexicostatistical methods on the comparative method.
- Section 5: Review of tree-building methods in lexicostatistics.
- Section 6: Review of the use of simulations of language change.
- Section 7: The use of n-grams for language identification and its extension to language comparison.
- Section 8: Detailed description of my proposed method.
- Section 9: The results of three sets of experiments:
    - Using artificial language data generated using simulated evolution.
    - Using a group of 10 Slavic languages.
    - Using 29 native Brazilian languages.

1.3 Acknowledgements

I would like to thank Bert Vaux, my supervisor, without whom this thesis would not have happened. I am grateful to Tandy Warnow who gave me an invaluable insight into the methods she and Don Ringe use. I also thank Antje Heinrich for her help with statistical matters and Sarah Hawkins and Ted Briscoe for helping me to source information on n-gram comparison techniques.

I would also like to thank the people who helped me with devising rules for translating from orthography to phonemes: Elliott Lash, Barbara Berti, Vicki Hart, Latifa Sadoc, Una Dimitrijevic and Gethin Jones.

2 Abstract

Traditional lexicostatistical methods which use comparisons of word lists to determine language relatedness are limited in that they can only be applied after the comparative method and only using very short lists of basic vocabulary, in order to mitigate the effects of borrowing. This renders them useless for application to languages that are not already well understood, and limits the statistical significance of their results. In this thesis, two variants of a new method based on n-gram comparison are proposed which eliminate these problems, and provide a means of automating the comparison of large numbers of languages on the basis of large volumes of text. The new methods can be applied before the comparative method, creating a much-needed tool for directing the energies of historical linguists. Additionally, a detailed simulation of sound-change is described, which is used to generate artificial languages, enabling the accuracy of lexicostatistical methods to be measured objectively.

3 Lexicostatistics

Although statistical techniques had been used to assess the relatedness of languages in the 19th century (see Hymes 1971), lexicostatistics as it is currently known was devised in 1950 by Morris Swadesh, in an attempt to determine the degree of relatedness of a number of North American languages. During the 1950s and 1960s lexicostatistics was used primarily for subgrouping of language families and for dating divergence of related languages (glottochronology); (see for example Swadesh 1954; Hirsch 1954; Baumhoff & Olmsted 1963).

By the 1960s, the assumptions behind lexicostatistics had been widely questioned and glottochronology had been discredited. Bergsland and Vogt's influential paper (1962) effectively ended the debate by disproving one of glottochronology's core assumptions—a common rate of change across languages. Indeed, glottochronology's three assumptions (that there is a common rate of change across languages, across time and across language features) are all demonstrably incorrect (see Coppin 2008:2-5).

Since the late 1990s, there has been a resurgence of work in subgrouping languages based on phylogenetic techniques, making use of techniques and ideas from genetics (see section 5.2). This adds rigour to the method, but has ignored, in many cases, the failings in the data being analysed and the assumptions on which the analysis is based (see section 3.2). It also relies on the soundness of the genetic methods themselves, and, more dangerously, on the assumption that genetics can be applied to language families at all (see section 5.3).

The usual lexicostatistical method involves the following steps:

1) Select a set of language varieties to be compared.
2) Select a meaning list—usually one of the lists devised by Swadesh (1950, 1955) with 100 or 200 meanings. (See Appendix A).
3) Collect a word list for each language being compared, with one word per language for each item in the meaning list.

4) Make cognate decisions between each pair of languages for each item in the meaning list.
5) Calculate the percentage of meanings which are cognate between the languages.

Once cognacy scores are calculated, additional steps are usually applied such as sub-grouping or dating on the basis of those scores.

3.1 Borrowing

Borrowing is often perceived as the greatest danger to the lexicostatistical method (see, for example, Black 2007, Dyen, Kruskal & Black 1992:30). Most lexicostatistical methods explicitly exclude borrowed forms, in the belief that loan-words would give an inflated view of the relatedness of two languages (see, for example, Kessler 2001:103-114).

As an example of the potential impact of borrowing, consider a lexicostatistical study comparing English and French. If all words were treated as potential cognates, regardless of whether they were known to be loanwords or not, the languages would appear to be extremely closely related. This is, in fact, an indication not that they are closely related genetically, but as Kessler (2001:109) puts it, that they are closely related historically. Hence, if lexicostatistics is carried out without dealing with loanwords, the results do not indicate degree of genetic relatedness (and thus, perhaps, should not be used for subgrouping or dating) but do provide information about historical relatedness, or surface similarity. In fact, as this thesis shows, it is possible to achieve quite accurate sub-grouping results using n-gram comparison without eliminating borrowing.

Swadesh (1950:159-160) took the rather cavalier view that "if one of the two languages displaces a word of the original stock, the second language may imitate the displacement or it may eventually cause the first language to return to the original form. Since these influences may be either in the direction of promoting or of retarding change, the trends may cancel each other out. The total drift: percentage of change may be the same as in the case of a single language out of contact with related languages, but the two languages will tend to stay together."

This assumption is clearly dubious, as it there is no reason to suppose that this drift should be equally balanced between the two directions. For example, Kessler (2001:106-107) examined the Swadesh 200 list for English and French and found 6 borrowings from Germanic into French (3% of the list) but only one borrowing from French into German (0.5%). Clearly if borrowing was ignored, and the drift assumed to be equally balanced between the two directions when comparing French and German, a significant error would be introduced.

3.2 Word lists

The correct length of list (of lexical characters) to be used in lexicostatistics has received a great deal of attention. In 1955, Swadesh shortened his 200 meaning list to 100 items. He explained that he had hoped to lengthen the list to increase the statistical accuracy of lexicostatistics, but that he could not find more than "a handful of really sound new items . . . while on the other hand defects in the old list were repeatedly made evident". He goes on to acknowledge that "quality is at least as important as quantity" and that "[e]ven the new list has defects, but they are relatively mild and few in number." (Swadesh 1955:124). Swadesh felt, and it has been almost unquestioningly accepted, that his list of 100 meanings was more "universal" than the 200 item list, and that it would therefore provide more accurate results.

Teeter (1963) argued that the degree to which a list of meanings is resistant to borrowing (and therefore its likely effectiveness for lexicostatistics) is inversely proportional to its length, leading to the conclusion that the "perfect list" would contain "no items at all". In contrast with Teeter, Guy (1980:37) felt that the lists should be "as long as possible", with 200 items as a minimum.

There exists, then, a fundamental tension for most traditional lexicostatistical methods: as they are dependent on basic vocabulary, their lists must be short (as there simply aren't enough basic words). On the other hand, the shorter the list, the less statistically significant the results will be. A method that avoids this difficulty by working with large volumes of text is described in section 8 of this thesis.

3.2.1 Garbage in: garbage out

The poor quality of the data used in many lexicostatistical studies combined with the shaky assumptions on which the work is based do not inspire confidence in the results, particularly given that much of the analysis is done on languages whose relatedness and histories are already well understood, and that when the lexicostatistical results differ from the accepted position, the researchers tend to provide explanations, rather than attempt to fix the methodology.

A recent study by Gray and Atkinson (2003) used statistical techniques combined with phylogenetic analysis to attempt to date the divergence of 87 Indo European languages. The study generated a great deal of interest because it appeared to confirm the theory that Indo-European "expanded with the spread of agriculture from Anatolia around 8,000-9,500 BP." However, Poser (2004) stated regarding their use of solely lexical characters that "[i]t's a little hard to believe that something as peripheral and unsystematic as lexical replacement provides sufficient information not only to reconstruct a realistic family tree but to date the splits." Furthermore, their study made use of the Swadesh 200 item list, which Swadesh himself already considered to be inadequate in 1955.

Swadesh's 200 item list appears to have become the standard list for most recent lexicostatistical work (McMahon & McMahon 2005; Ellison and Kirby 2006; Bryant 2006; Pagel, Atkinson & Meade 2007). It seems likely that this practice is due to the perceived value of the additional data available in the longer list, but ignores the dangers identified by Swadesh and others in using words that are not resistant to borrowing.

3.3 Basic vocabulary

As Gudschinsky (1956) explained, Swadesh was assuming that "some parts of the vocabulary of any language are [...] much less subject to change than other parts". It was on this basis that he devised his list of "basic" or "culture independent" words (which might be more accurately termed meanings, semantic slots or in the phylogenetic tradition: lexical characters). His intention was to select a set of meanings which were unlikely to be borrowed and likely to change at a relatively constant rate.

Swadesh (1955) showed that his 100 item list is more resistant to replacement than his original 200 item list, but it still contains 15 meanings which, according to his experimental findings, tend to be replaced at a rate of greater than 50% per 1,000 years.

Even the most basic words are subject to replacement. Borland (1982) conducted a set of lexicostatistical experiments using the Swadesh meaning lists (100 and 200) and also two longer lists made up of meanings randomly selected and believed to be susceptible to borrowing. The results were the same, within a reasonably small margin of error, showing that the basic vocabulary is just as susceptible to borrowing as any other randomly selected list of vocabulary.

Kessler (2001, pp 103-115) identified borrowings in a number of Swadesh lists. For example, he lists 41 borrowings for Albanian in the 200 item list and 16 in the 100 item list. For this reason, traditional lexicostatistics can only be applied after the comparative method (see section 4)—since even the most apparently basic list of meanings is subject to borrowing (16% for the Albanian 100 word list), reliable cognacy judgments are necessary in order to eliminate loanwords.

3.4 Application to non-lexical data

Lexicostatistical methods can be applied to aspects of language other than vocabulary, although such work is rare. Meillet (1925, 1970:48) pointed out that vocabulary is "the most unstable element of all in language", but explained that "in spite of this frequent instability of vocabulary it is the agreements in vocabulary which are immediately striking when languages are compared to each other." It appears that the preference for lexical characters is based, at least in part, on convenience: due to the large amount of work done using Swadesh's lists, data is relatively easy to obtain.

Swadesh (1951:12) felt that comparing vocabulary is "just as reliable" as comparing morphology, but with the additional benefit "that it can be converted into percentages with consequent advantages in objectivity". In contrast to this view, Teeter (1963:648) felt that the genetic history of a language is only recoverable using methods that take

account of "restructurings of the grammar" that have taken place. He felt that lexical comparison could be useful only as a first step in establishing genetic relationship as it "provides no way of going beyond lexical similarity".

Although Meillet considered morphosyntactic comparison to be essential to identifying linguistic relationship (Kessler 2001:95), Forster and Toth (2003) took the view that characters based on morphology and phonology, while usable for determining relatedness, were less reliable for constructing trees or for dating. Morphosyntactic characters are appealing for language comparison because it is believed that they are not often borrowed (see, for example, Ringe, Warnow and Taylor 2002:62). Kessler (2001:97) points out that in fact such borrowings do occur, and it is unsafe to assume that any commonalities are due to shared innovation. Kessler (2001:101) also points out that morphosyntactic characters can be hard to use because it can be difficult to know exactly what to compare—one character may not occur at all in a language, or might be conflated with other characters. Kessler's main objection to the use of morphosyntactic characters lies in the difficulty of devising a list of such characters in an unbiased way that will work with any language, rather than devising such a list on the basis of knowledge of the languages being studied (which could, of course, lead to experimenter's bias).

Dunn, Terrill, Reesink, Foley and Levinson (2005) applied phylogenetic methods to a number of Austronesian languages. They used 11 binary phonological characters (such as presence or absence of fricatives) and 114 binary morphosyntactic characters (such as article-noun order, pronoun number and presence or absence of suffix-marked possession). The tree they generated by applying these characters to 15 Papuan languages showed "a remarkably geographically consistent pattern". The authors admit that this is just as likely to represent the results of borrowing as genetic relatedness, although as the result was obtained from a set of languages for which the lexical data reveals no evidence of relatedness, the authors conclude that phonological and morphosyntactic structures may well provide access to greater time-depths than those available through lexical comparison.

A great deal of work has been carried out at the University of Trieste making use of syntactic parameters as characters for determining language relatedness (for example, Longobardi 2003, Rigon 2007). Longobardi (2003) argues that comparing syntactic parameters provides more reliable results than lexical comparison for long-distance relationships. This certainly seems reasonable, given the relative resistance of syntactic parameters to borrowing (Rigon 2007 examines the possible effects of parametric borrowing and finds evidence that it is less likely than parallel development). The problem with the parametric approach is that there is no generally agreed list of parameters (although Longobardi and his colleagues have a list of parameters that they have been using for some years now).

If a universal set of parameters can be agreed upon then analysis such as that carried out by Longobardi et al. is likely to be of great value, as it has much greater potential for working with a genetic (rather than phenetic; see section 0) model of language, and thus avoiding some of the problems inherent in traditional lexicostatistics.

3.5 Inspection methods

Ross (1950) proposed a method that involves the statistical comparison of correspondences between initial consonants. The main advantage of such inspection methods is that they do not rely on cognacy judgments—they can be carried out simply on the basis of inspection of the text of a language. In theory, this provides an additional advantage: inspection methods can be automated.

Inspection methods are potentially susceptible to similarities introduced by universals (such as onomatopoeia) or chance. Ross's method, like those of Ringe (1992) and Kessler (2001, 2007), used statistical methods to attempt to determine whether a detected relationship between two languages was likely to have occurred by chance.
Ringe's and Kessler's methods are useful for assessing the likelihood that a given pair of languages is related, but are not designed to analyse large numbers of languages in a pair-wise fashion. Additionally, Ringe's method is flawed in two ways (Kessler 2001:43-48). First, it uses an incorrect mathematical distribution as a model for significance (he uses

the binomial distribution, and Kessler and others have pointed out that the vastly more complex hypergeometric distribution is correct). More importantly, there is no real method for determining, based on Ringe's method, how likely it is that any given pair of languages are related, or to compare the likelihood of relatedness of two pairs of languages other than by fairly subjective measures.

Kessler (2001) describes a number of variants on Ringe's method, none of which produce particularly satisfying results in the tests he details. One problem with Kessler's methodology is that it relies, to some extent, on knowledge of the languages in question: he advocates removing loan-words from the word-lists, and also recommends removing any word that may have a common origin with another word in the list. These steps are clearly possible in situations where the level of knowledge of the languages and their history is good, but for cases such as the test on Brazilian languages described in section 9.5, they may not be possible.

Heggarty (2000) and Kessler (2001, 2007) have proposed methods that involve comparing phonemes on the basis of features such as voicing, place of articulation and nasality. Heggarty's method assigns varying weights to features (McMahon and McMahon 2005:214-219). He treats, for example, voicing as being less important than place of articulation, thus deeming /t/ and /d/ to be more similar to each other than are /p/ and /k/.

Further, Heggarty solves a problem that Kessler raised—how to decide which phonemes to compare—by matching forms through a template which consists of the reconstructed proto-form of the word. For example, Figure 1 shows how he compares the Italian *castello* /kastɛllo/ with the French *chateau* /ʃato/ using the Latin form *castellum* /kastɛllʊm/ as a template to decide which phonemes to compare.

Italian	k	a	s	t	ɛ	ll	o	
Proto-form (Latin)	k	a	s	t	ɛ	ll	ʊ	m
French	ʃ	a		t		o		
Comparison	k<=>ʃ	a<=>a		t<=>t		ll<=>o		

Figure 1: Illustration of Heggarty's method

Heggarty (2000:535) claimed that lexical meanings are "inherently unsuitable for quantification", because they provide data that is neither objective (because it is based on subjective assessment of cognacy) nor detailed (because it ignores the degree to which two words are similar). While it is clear that Heggarty's phonetic matching process is more objective than a cognacy-base approach, it is also clear that the process cannot be entirely objective, as the decision regarding how to apply the template is based on knowledge of the results of the comparative method. For example, Heggarty does not make clear why the /o/ in the French form is matched with the /ll/ in the Italian. It does not therefore seem likely that Heggarty's method could be fully automated such that a computer could apply it without being given explicit instructions regarding how to apply each proto-form as a template.

The method proposed in section 8 of this thesis is an inspection method that compares phonemes using n-gram analysis. This method meets both of Heggarty's criteria: it is objective and detailed. Like Ringe's method, it does not require cognacy judgments, but unlike other inspection methods it is able to work with large quantities of textual data, and can be fully automated. One of the main problems that this method solves is the reliance of lexicostatistical methods on the comparative method.

4 The Comparative Method

Swadesh (1953) claims that "lexical statistics may be used to help demonstrate a genetic relationship and need not be reserved only for use as a post-reconstructional exercise." In fact, proponents of lexicostatistics are often at great pains to make clear that they use the method not as a replacement for Meillet's comparative method, but rather as a supplement to it. For example, in McMahon & McMahon's (2005) preface, they reassure the reader:

> "What we are not doing [...] is trying to replace current historical-linguistic methodology with computer programs [...]. What we are suggesting is that it would be good for historical linguists [...] to incorporate some testing, simulation, and computational model-building in their work, in a way which has proved productive and interesting in corpus linguistics and sociolinguistics".

The reason for this caution is that the majority of lexicostatistical methods require knowledge of cognacy between meaning lists in the languages being compared. Ringe's (1992) approach was relatively rare in this regard in that it effectively automated the comparative method by looking for statistically unlikely correspondences between initial letters. Conversely, the majority of methods used by Kessler (2001) and McMahon and McMahon (2005) are reliant on accurate cognacy judgements. Heggarty's method, described in McMahon and McMahon (2005: 214-224), relies not just on cognacy judgements but also on reconstructed forms in a proto-language (or known ancestor cognates where applicable).

The fact that lexicostatistical methods can only be used after rigorous application of the comparative method is a weakness. It means that they are, at best, used to confirm relationships that are already well understood. While this can certainly be of interest, it does not appear to provide as much real worth as would be gained from applying the method to languages whose relationships were not already well understood.

Indeed, Teeter (1965) viewed lexicostatistics not as a method to be applied after the comparative method but as "a ground-clearing operation prior to historical research". He laments the "overinterpretation [by most lexicostatisticians] of the results of the ground-clearing as reflecting actual history". Hence, lexicostatistics, in Teeter's view, should be a method for helping to analyse a large set of data and thus to direct more manual research efforts. This approach is very rarely taken, but is the one of the key ideas behind the method proposed in this thesis (see section 8.1).

Dyen, Kruskal and Black (1992:18) make clear that they view the value of lexicostatistics as being solely in sub-grouping established families, and explicitly refute its use for establishing new relationships because of the lack of statistical evidence provided by comparing such small lists of words.

The next section of this thesis looks at the ways phylogenetic methods can be used to build language family trees; a technique that is usually dependent on the comparative method.

5 Building Trees

Prior to 1997, most work building trees from linguistic data made use of distance based methods such as the pair-group method (for example: Dyen, Kruskal & Black 1992:118).

5.1 The pair-group method

Given pair-wise relatedness measures for a set of languages, the pair-group method is applied as follows:

1. Combine the pair of languages with the highest similarity score into a sub-group.
2. Calculate a similarity score between this sub-group and each remaining language (or sub-group).
3. Repeat from step 1 until all languages are combined into a single tree.

The resulting tree is always binary (in other words, each non-terminal node has exactly two sub-nodes) which may not always be the most accurate way to represent a given language family, although Hale (2007:238) takes the view that "all changes introduce bifurcations into the descent tree" and that thus all language trees should be binary.

5.2 Phylogenetic methods

Most lexicostatistical work carried out since 1997 has used methods borrowed from genetics. This does, of course, rely on the validity of the genetic methods, and makes the unproven assumption that they are applicable to language relatedness. It also introduces practical difficulties in that most linguists are not qualified to determine the validity of genetic methods or their application to languages.

Methods such as those used by Ringe, Warnow and Taylor (2002) are character based methods in which trees are built up on the basis of character states. Characters can be lexical (i.e. whether or not a given word-meaning is cognate between two languages), phonological (presence or absence of a given phonological rule or innovation) or morphological (similarities between morphosyntactic features such as the form of the imperfect subjunctive). Methods such as UPGMA (in which the distance for a pair of

groups is considered to be the arithmetic mean of the distances between each pair of languages within the groups) and the pair-group method are distance-based, as they rely entirely on sets of distances between languages. Felsenstein (2004:147) explains that although distance-based methods appear to be less likely to produce reliable results than character-based methods, "the amount of information about the phylogeny that is lost in doing this [distance-based methods] is remarkably small. The estimates of the phylogeny are quite accurate".

The method proposed in this thesis is inherently distance-based, and the trees are generated using the pair-group method.

5.3 Waves and trees

It is not necessarily safe to assume that diachronic language relationships can be modeled accurately using a tree structure (Stammbaum). In this model, languages are considered to be related to each other in much the same way that species of animals are (Schleicher, 1863); it contrasts with Schmidt's (1872) wave model which models linguistic innovations as waves, spreading independently from one dialect to the next. Teeter (1965:1522) took the view that a strict tree model of language change is not accurate, and thus claimed that lexicostatistical dating is fatally flawed.

Sankoff & Sankoff's (1976) study attempted to show whether the tree model or the wave model better fitted the facts of a 26 Papua New Guinean languages. Their finding was that the tree model made a better fit once allowance was made for borrowing (see section 6).

The method proposed in this thesis extends Sankoff & Sankoff's idea by combining the tree model with a limited version of the wave model (see section 8.4).

5.4 Phenetics or genetics?

A major difficulty with all lexicostatistical methods that are aimed at phylogenetic tree generation is that, in fact, the work is not "genetic" in nature at all. Phylogenetics originated in the biological sciences, and was a replacement for phenetic study, which involved examining the physical differences between creatures (their phenotypes, as opposed to their genotypes) and using those to attempt to build genetic trees. The main problem with phenetics is that it can be easily misled by phenomena such as parallel or convergent evolution. For example, wings appear to have evolved independently in birds and insects (Nichols 2006). Similarly, common properties can emerge in unrelated languages, such as apparent linguistic universals. Borrowing between languages can have a similar effect.

Phylogenetic study in biology largely avoids this problem by examining the DNA structure (the genotype), rather than the phenotypic realization of the genotype. Linguists have attempted to replicate this by looking at, for example, cognacy judgments across a fixed set of "basic" words. Unfortunately, this is still closer to phenetics than it is to genetics—Teeter (1963:641) points out that "the lexicon is nothing but the outward face a language turns to its associated culture." The method proposed in this thesis attempts to mirror genetic analysis by comparing statistical models of languages' underlying phonotactic rules.

6 Simulation of Language Change

A small body of work has been growing in recent years that assumes that lexicostatistics itself is valid, and tests specific lexicostatistical methods using simulated language change. To date, all such work has been based on extremely simplistic models of language. One of the first such pieces of work was carried out by Guy (1980), comparing a set of 7 algorithms. His experiment involved creating a set of 19 invented languages and running a computer program which simulated evolution of the languages over a period of time, on the basis of a tree defined by Guy. The results of that simulation could then be compared with the original tree to determine their accuracy.

Barbançon, Warnow, Evans, Ringe and Nakhleh (2007) attempted similar simulations testing phylogenetic methods. Their results showed that the standard distance-based methods (as used by most lexicostatistical studies before 1997) are out-performed by character-based methods such as maximum parsimony (a phylogenetic technique that attempts to builds trees which involve the fewest changes of state). This comparison was based on comparison of cognacy of lexical items, presence of phonological rules and morphosyntactic characters.

Sankoff & Sankoff (1976) carried out an experiment using lexicostatistics to generate a phylogenetic tree, and also designed a simulation of the wave model. Their aim was to determine which model provided the best fit to the observed facts. In following their tree approach, they assumed that a perfect phylogeny (see Felsenstein 2004:95) would result if the tree model was a good representation of language relationship. This is valid reasoning, although it does assume that there is no undetected borrowing, and given that they were examining languages whose relationships were not well understood, this seems a risky assumption. For their wave model, they used a multidimensional scaling technique which plots the languages as areas in a two dimensional space, intending to illustrate the real distance between the languages, taking into account features such as mountains. Based on observed retention rates, they then built a model that allowed randomly introduced innovations to spread among the languages along a wave front.

Their simulation simply simulated the transmission of lexical innovations using a statistical model. It took no account of linguistic factors, and represented a purely abstract, statistical view of the way that languages change. Sankoff & Sankoff also admitted that their wave model simulation assumed that each speech community was static, and that "more realistic versions would have to allow for migration or movement of speech communities in the course of simulation" (1976:35). Sankoff & Sankoff concluded that the best model would be one that combined features of both tree and wave.

Embleton (1986) combined the wave model and tree model in a single simulation. As with Sankoff & Sankoff's model, Embleton's was purely based on lexical comparisons. Sound change is not relevant to these models, as they are only interested in rates of replacement of cognate forms. This, of course, assumes that borrowing is always detectable and that cognacy can always be accurately judged.

6.1 Simulating sound change

Hartman (2003) developed a program (Phono) for simulating sound change on the basis of a set of prescribed (regular) sound change rules. His program is, in some ways, more sophisticated than the system proposed in this thesis (for example, its notation for specifying sound change rules allows more complex contexts to be specified). However, it differs in two important ways:

1) It was developed simply to simulate sound change, not to test lexicostatistical methods.
2) It allows any sound change to take place (for example, a > tʃu / #__) without regard for the frequency of attestation of types of changes or of resulting phonemic inventories.

Neither of these is intended as a criticism of Hartman's system; I merely observe that they are fundamental differences between his system and mine.

7 n-gram Comparison

My system, described in section 8, uses n-gram comparison as a mechanism for measuring language similarity. An n-gram is a set of n consecutive characters (when n is 3, the resulting n-grams are usually referred to as trigrams).

N-grams have been traditionally used for language identification (Damashek 1995). The usual method for identifying a language from a piece of text is to calculate an n-gram vector for the text, and compare it (using standard vector comparison methods, as described in section 7.1) to the vectors for each known language. The closest vector is taken to indicate the correct language for that text.

An n-gram vector for English might show that the trigrams *ing* and *ted* are extremely common, while *xjq* never occurs. Each slot in the vector represents the frequency of single n-gram. For trigrams, the vector contains 26x26x26 = 17,576 items (assuming the English version of the Roman alphabet). A 1-gram would simply represent the relative frequencies of each letter in a language.

A partial hypothetical trigram vector for English is shown in Figure 2.

aaa	aab	aac	...	ing	inh	ini	...	tec	ted	tee	...	zzx	zzy	zzz
0	0	0	...	0.1	0	0.001	...	0.002	0.09	0.005	...	0	0	0

Figure 2: Partial trigram vector for English

Typically, n-grams include spaces as well as letters. The intention behind this is to encode word boundaries. Hence, the word "fish" might be encoded as the following trigrams: ␣fi, fis, ish, sh␣. (Note that ␣ is used to represent a space).

7.1 Using the dot-product to compare n-gram vectors

A vital aspect of an n-gram based system is the method that is used to compare n-gram vectors. The standard approach is to use the normalised dot-product, defined as follows:

$$S_{mn} = \frac{\sum_{j=1}^{J} x_{mj} x_{nj}}{\sqrt{\sum_{j=1}^{J} x_{mj}^2 \sum_{j=1}^{J} x_{nj}^2}}$$

S_{mn} is the dot product of the vectors for languages m and n.

J is the total number of possible n-grams.

x_{mj} is the relative frequency of occurrence of n-gram j in the text for language m. Note that the sum of all x_{mj} for a given m is 1.

Before this comparison can be applied, the vectors are normalised, meaning that all of the values in each vector are reduced proportionally until the sum of the values in the vector is 1. This ensures that the vectors are comparable.

Damashek's (1995) Acquaintance algorithm uses n-grams to identify the subject matter of documents. He notes that the dot product "can provide a gross measure of similarity—in particular, language discrimination is excellent". It is this ability to discriminate languages that is the basis of my method. Traditional lexicostatistical techniques effectively compare languages by comparing a selected subset of vocabulary items. N-gram comparison provides a way of identifying the extent to which texts in two languages differ, but the same comparison can be used to determine the degree of similarity (and thus, to a limited extent, the degree of relatedness) between two languages.

7.2 N-grams for phylogenetic analysis

Scannell (2004) applied n-gram comparison to texts automatically gathered from the internet for 425 languages. His method applied n-grams based entirely on orthography, meaning that two languages with different alphabets would always be deemed entirely dissimilar. Hence, for example, his method would deem Serbian and Croatian to be as different as Japanese and English, because although Serbian and Croatian are mutually intelligible, they use different writing systems. He applied a standard phylogenetic method (neighbour joining) to his data and generated a tree which, in spite of the obvious weakness of the method, was reasonably accurate at the very coarse-grained group level. It is clear that converting texts to IPA phonetic notation, as in my method (described in section 8), provides an improvement over the performance of Scannell's system.

Huffman (1998) applied n-gram comparison to the orthographic representation of texts. He applied the method to a group of European languages, and to a group of native American languages. Huffman used Damashek's Acquaintance algorithm which primarily differs from the approach I took by computing a centroid vector, the average of all the n-gram vectors for the documents being examined. This provides a representation of the common features amongst the languages. This centroid vector is subtracted from the language vectors, as a way of eliminating common elements across languages. Huffman (1998:216) concluded that "the distribution of sound patterns, even when they are poorly reflected through alphabetic representations, is a fairly reliable marker of the genetic relationship among languages".

N-gram analysis has also been used in biological phylogenetics to carry out sub-grouping of animal species. Stuart, Moffett & Baker (2002) used n-grams of peptides and protein sequences in genome sequences to produce phylogenetic trees of a number of species of mammals. Stuart et al. used vector dot-products to produce similarity matrices between pairs of species, and then used UPGMA (a variant of the pair-group method used here) to produce trees. Their trees were produced using various configuration settings, and each tree was evaluated against known data to determine its accuracy.

8 My Solution

This section proposes a new lexicostatistical method, as well as a sophisticated model of sound-change designed to facilitate rigorous statistical testing of the method.

8.1 Motivation

Since its creation, lexicostatistics has suffered from a number of serious flaws (see sections 3.1, 3.2, 3.3 and 4). Many of these flaws relate specifically to dating, but its use for determining genetic relatedness, for sub-grouping and even for simply determining historical relatedness is also problematic. This thesis proposes a new method which does not address all of these problems, but is designed to address some of the most fundamental. This section contains a summary of the motivation behind my proposed method, as well as an indication of its likely limitations.

1. **Simulation of language change.** This is used to create a set of artificial languages whose relatedness is known perfectly. This ensures that the comparison method is objectively testable. Most lexicostatistical studies compare their results against established (but hypothesised) views of relatedness. This approach is limited in that it is only as accurate as the model tree. Additionally, it means that such methods must always be applied to languages whose relatedness is well understood. It is hoped that by simulating language change, the parameters of the system can be better understood, and a degree of confidence can be gained that the method works. That simulation of this kind is not widely used is a great weakness of lexicostatistics: the usual method is to produce a tree and simply compare it with what is generally accepted. If the tree is close enough and any differences can be explained then the method is deemed a success. This is not a particularly scientific approach, and I believe that the use of simulation would greatly enhance the reputation and scientific rigour of the field.
2. **Trees and waves**. The assumption that languages change in a tree-like fashion is the basis of much historical linguistic work. The wave model, in contrast, is opaque and does not lend itself to lexicostatistical analysis. The present model is

largely tree-based, but the simulation attempts to incorporate some aspects of the wave model by allowing borrowing between languages. The reason for basing the system on the tree model is that it provides a direct metric for measuring success: a tree can be produced and compared with a known model tree. It is hard to envisage a similar comparison method for data generated using a pure wave model.

3. **Sub-grouping and dating.** Dating on the basis of lexicostatistics is extremely suspect. On the other hand, sub-grouping, while subject to some of the issues associated with dating, is more tenable. My method, although strictly speaking determining historical rather than genetic relatedness, attempts to reproduce genetic trees for languages. One of the main reasons for this is that it provides a mechanism for objectively determining the effectiveness of the method (comparing the trees produced with model trees whose accuracy is known).
4. **N-grams.** N-grams are routinely used for language identification. N-gram analysis is based on comparing statistical models of languages, and as such is closer to a genetic analogy than comparing lexical items. N-gram comparison can also be applied to very large volumes of text and can be automated.
5. **Borrowing**. My method does not attempt to eliminate or otherwise deal with the effects of borrowing. This is certainly a limitation if the method is being used for sub-grouping (although the results obtained suggest that borrowing is not as serious a threat to the method as might be supposed) but is not relevant when the method is being used to determine historical relatedness. This removes one of the main limitations of traditional lexicostatistics methods which can only be applied to very short lists of (supposedly) basic vocabulary.
6. **Volume of text**. Due to the use of n-gram comparison and the primary aim of assessing historical relatedness rather than genetic relatedness, very large volumes of text can be analysed using the present method. This will be shown to provide greater statistical significance than can be obtained using very small amounts of text.
7. **Comparative method.** Almost all lexicostatistical methods are dependent on cognacy judgments determined by the comparative method. This is a great

weakness, as it means that they can only be applied to languages that are already well understood. My method eliminates this requirement by using the statistical n-gram model of language, thus meaning that the method can be applied to any set of languages. It is ideally suited to being applied before the comparative method, pointing to languages or families whose potential relatedness is most worthy of further study.

To summarise, the main benefits of my method are that it can be applied in an automated fashion, using large amounts of text, and can be used to direct energies on languages that are not well understood, prior to application of the comparative method. The main weakness is that it does not discount borrowings, and thus is really only able to determine historical relatedness, rather than genetic relatedness. However, it is clear that historical relatedness and genetic relatedness are, in many cases, similar concepts, and the results below indicate that this is not as great a weakness as it might appear.

8.2 The simulation model

My system uses a detailed simulation of sound change which is based largely on the tree model but also incorporates aspects of the wave model. The algorithm for evolving a single language into multiple descendants is as follows:

1. Select a random phoneme from the language's current phonemic inventory.
2. Identify a possible change (conditioned or unconditioned) that could be applied to that phoneme.
3. Decide whether to add that rule to the language's phonology.
4. Repeat from step 1 for a pre-determined number of iterations.

The possible sound changes allowed by the system are listed in Table 1.

Type of change	**Example**
Voicing and devoicing of consonants	g > k
Vowel shifts	e > ɛ or i > ɨ
(De)palatalization of consonants	n > ɲ
(De)spirantization of stops	b > β
Nasal place assimilation	np > mp
Velarisation	n > ŋ
Glottalisation	t > ʔ
Epenthesis	mr > mbr, ns > nts or pr > pər
Cluster reduction	mbr > mr
Apocope	i > Ø / _#
Syncope	i > Ø / V_V
Aphaeresis	i > Ø / #_
Rhotacism	s > r
Lenition	p > f
Fortition	j > dʒ
Excrescence	t > tə / _#
Vowel breaking	i > iə
Affrication	t > tʃ
Gemination	p > pp
Degemination	pp > p
Compensatory lengthening	arta > aːta

Table 1: Types of sound change allowed by my system.

(These changes are based on the types of possible sound changes described in Campbell 1998:17-46, Crowley 1992:36-47 and Hock & Joseph 1996:126-134).

Each sound change can have a specified context (e.g., must be word-final or word-initial, must occur before a back vowel, must occur between two velar consonants) and a probability. The probability, for example, of lenition is much greater than that attached to fortition, reflecting the attested frequencies of these changes in the world's languages (Crowley 1992:38).

The changes are bound by a set of meta-principles which ensure that the resulting language is linguistically plausible. For example, meta-principles forbid diacritics to be used to create meaningless combinations such as w^w, j^j, or $\tilde{s}$.

The use of rules whose nature, context and probability are based on attested sound changes was intended to enable the system to simulate sound change as closely as possible to reality.

Once a suitable change has been identified, the decision to apply it depends on the following factors:

1) Is the target sound of the change already present in the inventory of the language? (Most changes prefer targets that are already present, as otherwise evolution over time has the effect of increasing the size of the phonemic inventory. Some changes, such as vowel shifts, prefer sounds that are not currently present in the phonemic inventory).
2) If the rule is a copy or reversal of a rule already present in the language's phonology then it is rejected.
3) If the target sound is not already present in the phonemic inventory, it is looked up in a table which contains each of the phonemes used in the system, along with the count from Maddieson (1984) of how many languages use that phoneme. A probabilistic choice is then made, ensuring that the less commonly occurring sounds are more likely to be rejected. Hence, for example, the change b > β is more likely to be accepted than b > в, as β occurs in 38 of the languages in Maddieson 1984, while в occurs in none.
4) Finally, a random element is introduced which may reject any change. This random element is proportional to the probability assigned to the change, ensuring that common changes occur more frequently than those assigned low probabilities.

For example: an application of the system to Spanish, allowing 250 iterations, added the following sound changes to the language's phonology:

1.	a > e	9.	g > ɣ
2.	o > u	10.	θ > f
3.	u > ue / r_	11.	s > r / V_V
4.	k > tʃ / _i	12.	i > Ø / C_#
5.	x > ɣ / #_	13.	s > ss / u_u
6.	eʌʌ > eeʌ	14.	m > β
7.	s > ʃ / r_	15.	u > iu / _b
8.	k > g / u_u	16.	iɛ > u

In these examples, C represents any consonant and V represents any vowel. Although the system does not use a feature-based representation for sounds, it is possible to specify types of sounds such as "front vowels" or "voiceless consonants" in the rules.

The changes generated by this algorithm are appended to the end of the language's phonological rules, and subsequent evolution can be applied to the augmented phonology to produce a more distant descendant.

It is important to note that this simulation of sound change assumes that sound change is entirely regular (the Neogrammarian hypothesis), and does not allow for lexically gradual changes. If the Neogrammarian hypothesis is correct, then the model is accurate. Alternatively, it could be assumed that the system is modeling the gradual process of sound-change in a binary manner by simply jumping from the start of the change to the point at which it completes.

8.3 Simulation example

This section contains an example of a simple execution of the evolution process, starting from a single language (Spanish), without borrowing, and producing a single offspring at the end of each generation. Three generations were run, producing Spanish, Spanish1, Spanish11 and finally, Spanish111. Each generation applied 250 iterations of the evolution algorithm—a large number, intended to result in a noticeable set of changes.

A short piece of text was used, for the sake of clarity:

Como en otros países, es común que en los estadios las "hinchadas" estén organizadas por grupos de adeptos denominados "barras bravas", que impulsan los cantos de apoyo a los equipos, y los viajes de simpatizantes cuando juegan de visitante.

(Source: http://es.wikipedia.org/wiki/Futbol_en_Argentina. Accessed 20th May 2008).

The IPA transcription of this text, according to the rules of the initial parent language (Spanish) is:

komo en otɾos paises es komun ke en los estaðjos las intʃaðas esten oɾɣaniθaðas poɾ gɾupos de aðeptos denominaðos baras bɾaβas ke impulsan los kantos de apoio a los ekipos i los bjaxes de simpatiθantes kwando xweɣan de bisitante

After one generation, the following new phonological rules were added:

ei > i	tʃ > f	r > br / m_
ʎ > ɲ	i > ɨ	ɨ > ɪ
g > Ø / a_s	t > ts / _a	ɦ > h
ɲ > m	m > Ø / ɨ_ ð	d > ð / _VC
t > tə / _t	n > ŋ / _#	u > uɪ / l_
s > x / ɾ_	ɣ > x	p > ɸ / _VC
e > ɦe / #_	o > u	b > Ø / _b
ie > u	p > Ø / _p	
s > r / V_V	u > hu / #_	

The resulting IPA transcription for this generation (Spanish1) is:

kumu heŋ hutɾus paɪɾes hes kumuŋ ke heŋ luɪs hestsaðjus las ɪnfaðas hesteŋ huɾxanɪθaðas puɾ gɾupus de aðeptus denumɪnaðus baras bɾaβas ke ɪmpulsaŋ luɪs kantus de apuɪu a luɪs hekɪpus ɪ luɪs bjaxes de sɪmpatɪθantes kwandu xwexaŋ de bɪɾɪtsante

After the second generation, the following new phonological rules were added:

h > ha / _#	p > b / ə_a	ð > ðɪ / _#
f > uf / l_	k > tʃ / _e	ɪ > i
h > ʔ	g > Ø / e_ ɪ	ʝ > ʝə / _#

The resulting IPA transcription for this generation (Spanish11) is:

kumu ʔeŋ ʔutrus paires ʔes kumuŋ tʃe ʔeŋ luis ʔestsaðjus las infaðas ʔesteŋ ʔurxaniθaðas pur grupus de aðeptus denuminaðus baras braβas tʃe impulsaŋ luis kantus de apuiu a luis ʔekipus i luis bjaxes de simpatiθantes kwandu xwexaŋ de biritsante

In the final generation, the following new phonological rules were added:

a > e	p > b	d > Ø / ə_j
x > g	s > ɾ / V_V	mg > ŋg
ð > f	f > v	k > kØ / _d
k > Ø / _k	b > ɖ	β > b / i_
r > s	i > ʔi / #_	n > m
ɾ > ɾɾ / i_i	θ > θə / _#	ɖ > Ø / _#
g > ɣ	ɣ > Ø / _ɣ	t > pt / m_

Following this final set of rules, the IPA transcription for Spanish111 is:

kumu ʔeŋ ʔutrus ɖeiɾes ʔes kumuŋ tʃe ʔeŋ luis ʔestsevjus les ʔimveves ʔesteŋ ʔurɣemiθeves ɖur ɣruɖus de eveɖtus demumimevus ɖeɾes ɖɾeβes tʃe ʔimɖulseŋ luis kemptus de eɖuiu e luis ʔekiɖus ʔi luis ɖjeɣes de simɖetiθemptes kwemdu ɣweɣeŋ de ɖiɾitsempte

8.4 Borrowing and the wave Model

A full simulation of the wave model of language change is virtually impossible: at the least, it would require a simulation of large numbers of individuals, each of which had their own language (Hale 2007), and a model of language transmission and diffusion of changes would need to be built. Some models of language change based on the wave model have been built (Sankoff & Sankoff 1976 and Embleton 1986) but these have modeled language diffusion in a very simplistic manner.

My system is based largely on a tree model, but incorporates lexical and phonological borrowing. Assuming that each language produces two direct descendants in each generation (i.e., the language tree is binary), the spread of languages in a simulated (simplistic) geographic space follows the pattern shown in Figure 3.

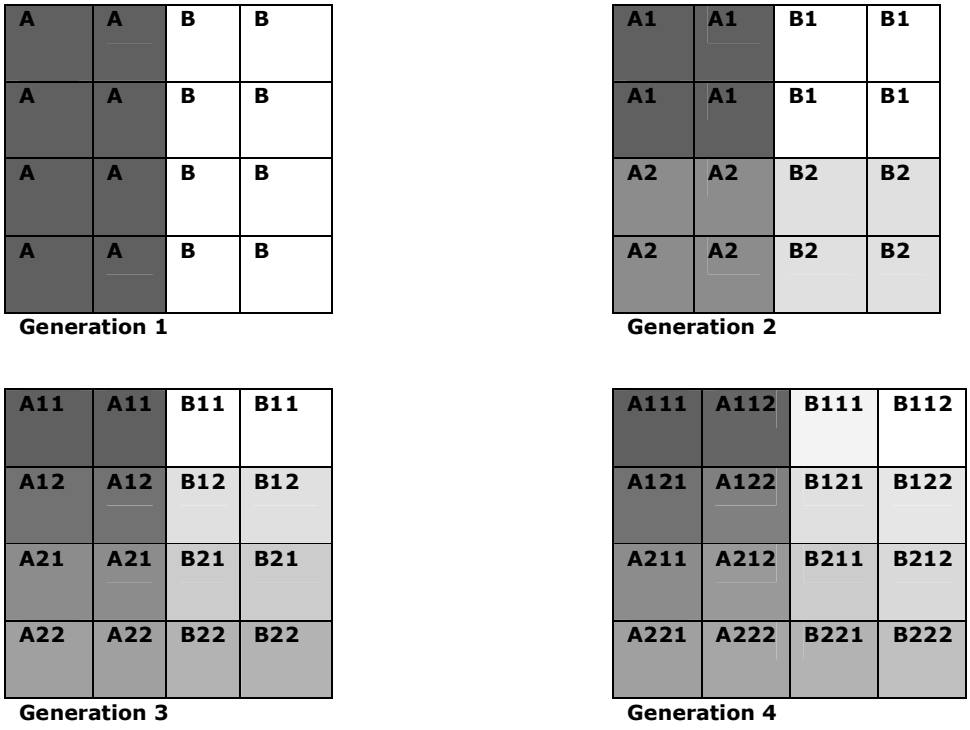

Figure 3: The spread of languages across a geographic space.

In the first generation, each language (A and B) occupies half of the available space. In the second generation each language splits into two (A1, A2 and B1, B2), splitting the parent language's space in half. This process repeats until, after the fourth generation, there are sixteen languages, each occupying one sixteenth of the space originally occupied by two languages.

In fact, the areas occupied are not relevant, only the relative distances between languages. These can be used by the system to determine the likelihood of borrowing. There are four configurations for the system:

1) No borrowing allowed.
2) Unrestricted borrowing allowed, regardless of distance.
3) Borrowing only allowed between adjacent languages.
4) Rate of borrowing between a pair of languages inversely proportional to the square of the distance between them.

In the third configuration, where borrowing is restricted to adjacent pairs, language A212 could borrow from (or be borrowed from by) A122, A211, A222 or B211. (Diagonal adjacency is not allowed). In a given generation using this system configuration, a language can borrow from any or all of its neighbours.

Hence, a sound change which is innovated by language A in generation 1 may be borrowed from language A1 to B1 in generation 2; inherited by B12 and then borrowed by B21 in generation 3; and inherited, finally, by B211 and B212 in generation 4.

Although this is not intended to be an accurate or complete simulation of the wave model of language change, it is intended to introduce an element of lexical and phonological diffusion to ensure that the system is not constrained to simulating only tree-based language change.

8.4.1 Lexical borrowing

After the addition of sound change rules, a language in this simulation can borrow a number of words from other languages (lexical borrowing), using repeated applications of the following process:

1) A native word is selected for replacement.
2) A word with a similar frequency of occurrence is selected from another language.
3) The borrowed word is converted into the borrowing language's alphabet by first converting to the Roman alphabet using a set of mappings (such as ц → ts, for Cyrillic) and then converting from the Roman alphabet to the borrowing language's orthography.
4) The text for the borrowing language is modified by replacing all occurrences of the native word with the borrowed word.

Hence, the effect of lexical borrowing by a language is purely to alter the contents of the text that is associated with that language.

The process of converting a borrowed word into a form that is appropriate (phonologically and orthographically) for the borrowing language is described by Hock and Joseph (1996:262-263) as nativization. They describe two main mechanisms for nativization: phonological and lexical. Phonological nativization involves mapping the sounds of the borrowed word to the borrowing language's phonology (and phonotactics), thus ignoring the orthographic form altogether. Lexical nativization involves borrowing the orthographic form, and also involves a certain amount of phonetic or phonological nativization to ensure that the word is pronounceable.

The system proposed in this thesis is effectively using a form of lexical nativization. Each word is borrowed in its orthographic form, and that form is converted to the closest match in the borrowing language's orthography. The pronunciation of the borrowed word is then determined by the borrowing language's phonological rules.

Table 2 shows four examples of nativization of borrowed forms generated by the system.

Donor language	**Borrowing language**	**Original orthography**	**Original phonemic form**	**Borrowed phonemic form**	**Borrowed orthography**
Spanish	Russian	militar	militaɾ	mʲiɬʲitar	милитар
Macedonian	Basque	читајќи	tʃitajci	tsitaxki	tsitajki
Georgian	Polish	ჯი	kʼi	kʲi	ki
Turkish	Swahili	iddialaşıyorlardı	iddialaʃɯjoɾlaɾdɯ	iɗɗiɑlɑsijɔrlɑrdi	iddialasiyorlardi

Table 2: Nativization examples.

For example, in borrowing the Georgian word "ჯი" into Polish, the system takes the steps shown in Table 3.

Step	Description	Current form
0	Georgian orthographic form	ჯი
1	Convert to IPA	kʼi
2	Convert to Roman script	ki
3	Convert to Polish alphabet (no change)	ki
4	Convert to IPA	kʲi

Table 3: Example of the way the system nativizes a Georgian word into Polish.

8.4.2 Phonological borrowing

Language A can borrow a phonological rule from language B using to the following procedure:

1) A rule from the phonology of B is selected whose left hand side is a sound currently present in the phonemic inventory of A.
2) The change is rejected if it is an exact copy of a rule already present in A's phonology, or if it is an exact reversal of a rule in A's phonology.
3) If any of the sounds on the right hand side of the rule is not currently in the phonemic inventory of A, then the data from Maddieson 1984 is applied (see section 8.2) to bias the system towards introducing common sounds.

4) If the rule is not rejected, it is added to the end of A's phonology (see section 8.7)

Any innovated rule in a language's phonology can be borrowed. The phonological rules of a language, in this system, consist of rules for converting orthography to phonemes, and rules for converting one phoneme into another. Only the latter kind can be borrowed.

8.5 The n-gram comparison method

The method described in this thesis is based on calculating an n-gram vector for a large corpus for each language being compared, and calculating vector distances between all pairs, which can be treated as similarity ratings. The n-gram vectors are built on phonetic symbols rather than the letters of the alphabet. Specifically, each n-gram consists of n IPA symbols. Hence, for example, a set of 4-grams might be:

ʃpað, afap, ðuʃa

There are more than 100 IPA symbols, resulting in a potential 4-gram vector of 100x100x100x100 = 100 million entries. In practice, n-gram vectors are extremely sparse, and so storing this data on a computer is entirely feasible.

In addition to the question of which n-gram comparison method to use, it was also necessary to decide which value of n—the n-gram window—to use. Trigrams are most commonly used (see, for example, Shannon 1948) presumably because trigram vectors are of a manageable size but contain more information than unigrams or bigrams. For this thesis, experiments were carried out with windows of 1, 2, 3, 4 and 5.

The method being applied in this thesis can be described, briefly, as follows:

1. Convert all text to IPA phoneme representation.
2. Calculate an n-gram vector for a piece of text in each language.
3. Create a distance matrix based on pair-wise comparisons of the n-gram vectors.
4. Attempt to build a phylogenetic tree showing genetic affiliations based on the distance matrix.

8.6 Pair-wise n-gram comparison

In addition to the standard n-gram comparison method described in section 8.5, a method was devised which, like the initial phoneme comparison methods of Kessler (2001) and Ringe (1992), involved pair-wise comparisons, but using n-gram comparison rather than comparing initial phonemes.

To compare a pair of languages, using a pair of Swadesh lists:

For each word, an extremely sparse n-gram vector is created. The two n-gram vectors for a given meaning are then compared. The average of the scores for all pairs of words is used as the similarity score. This contrasts with the original n-gram comparison method which compares n-gram vectors for each document, rather than for individual words. The difference is illustrated in Figure 4.

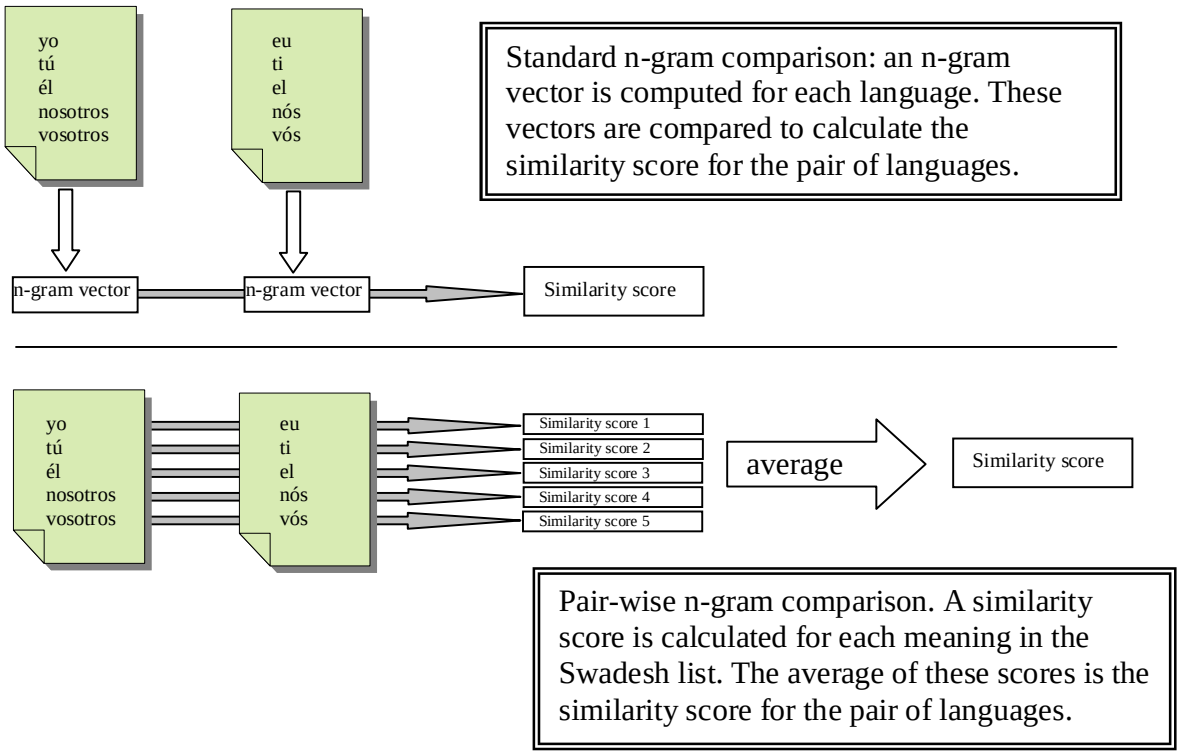

Figure 4: Illustration of the standard and pair-wise n-gram comparison methods.

8.7 Phonetics and phonology

In simulating language change it was important to consider *what* precisely would be changing. I decided to simulate and compare languages at a phonological level, rather than a phonetic level. This was not just important in relation to the simulation of change, but also was vital to the entire system, as it determined the data that would be analysed. For example, many Slavic languages have final consonant devoicing; the Russian word лёд is usually pronounced something like [l^{j}ot]. The underlying representation for this word, though, is /l^{j}od/, which is converted to the phonetic form by the application of a devoicing rule.

A difficulty with working on the phonetic level is the lack of consistency. Even languages that have a "standard" version have a great deal of variation between speakers. The phonological representation of the sounds of a language, however, appear to be relatively stable. Hence, a reasonable representation of the state of a language can be found in its phonology, whereas the phonetic level represents the differences between dialects or even individual speakers.

Hale (2007:101) reflects the generativist view of sound change that "while changes were often observed at the phonetic level, the primary mechanism of getting a different form at that level was to modify the phonological component". Bynon (1977:114) states this more explicitly: "according to Halle (1962) the basic type of sound change, namely innovatory change, is reflected in the transformational model by the addition of a new phonological rule at the end of the phonological component." This is the approach taken by my system.

Working at a phonological level has practical benefits: in Russian, for example, the pronunciation of vowels is unpredictable at a phonetic level without knowledge of stress patterns (unstressed non-high vowels are usually reduced). The system would not be able to deal with this variation, as stress is not predictable from the orthography. Hence, for example, my system can transcribe один as /od^{j}in/, although phonetically it is [ɑd^{j}in^{j}].

Another advantage of working at this level is that it allows the system to model underlying phonotactics rather than surface phonotactics (see, for example, Shibatani 1973 and Sommerstein 1974). I am taking the view that surface, phonetic level, phonotactics are equivalent to phenetics while the underlying, phonological phonotactics are closer to being analogous to genetics.

8.8 Converting text to phonemes

The method of comparison used by Scannell (2004) employed n-grams of graphemes. This has the advantage of being extremely simple to set up (the system simply works from written text) but has two main disadvantages which greatly outweigh this advantage:

1. It considers languages which have different alphabets—but are otherwise mutually intelligible—to be entirely dissimilar (e.g. Serbian and Croatian).
2. In working with languages such as English which have a deep orthography (explained below), the system is, in effect, analysing the somewhat arbitrary spelling rules chosen by the language.

My method avoids these problems by converting written text into phonemic representation using IPA symbols. This means, for example, that Armenian can be compared with Spanish, Ukrainian, all of which are written in different scripts.

Van den Bosch, Content, Daelmans and de Gelder (1994) define the orthographic depth of a writing system as "the degree to which it deviates from simple one-to-one letter-phoneme correspondences". Hence, for example, English has a deeper orthography than Spanish.

In order to convert text to phonemes, it was important to work with languages whose orthography is reasonably shallow. It is straightforward to build rules for languages with shallow orthography (such as Spanish), which express the regular correspondences between written letters and phonemes. For example, some simple rules for Spanish are shown in Table 4.

letter(s)	IPA notation
ch	tʃ
p	p
ñ	ɲ
rr	r

Table 4: Example of phonological rules for Spanish

Some rules are contextual, meaning they express the way a letter is pronounced depending on letters that come before or after it, as shown in Table 5.

letter(s)	IPA notation	Context
c	θ	before i or e
c	k	default
g	x	before e
g	ɣ	before a vowel or r or s
g	g	default

Table 5: Example of contextual phonological rules for Spanish.

The ordering of these rules is important. The system, in converting the letter "c" to phonemes, will examine the rules in order. If it finds a following "e", it will convert the letter to θ. Otherwise, it will apply the next matching rule.

8.9 Chunking

The method being employed here makes use of IPA transcription. This means that, for example, the vowels /ɛ/ and /e/ are considered to be different. In order to test whether this level of detail is necessary, and whether it is an effective way of comparing languages, my experiments were carried out using five chunking levels.

The effect of the five chunking levels on consonants is shown in Table 6.

Chunking Level	Consonants
1	Full IPA transcription
2	Diacritics ignored (for example, p^w becomes p)
3	No voicing distinction (for example, b becomes p and d becomes t)
4	No place of articulation distinction (for example, all plosives become t)
5	Nasals grouped with plosives Fricatives, approximants and laterals fricatives and approximants grouped together Trills, flaps and lateral flaps grouped together

Table 6: Effect of chunking levels on consonants.

The chunking of consonants is based on features. The chunking of vowels is, of necessity, somewhat more arbitrary, as can be seen in Table 7.

Chunking Level	Vowels
1	Full IPA transcription
2	Diacritics ignored (for example, the nasal vowel ã becomes the oral equivalent, a).
3	vowels grouped as: i: i, ɨ, ɪ, ɯ u: u, ʉ, ʏ, y, ʊ e: e ɛ: ɛ ə: ə, ɘ, ɵ, ɜ, ɞ, ɤ a: a, æ ɑ: ɑ, ɐ ɔ: ɔ, ɒ, ʌ ø: ø, œ, ɶ o: o
4	The groups defined in level 3 are further grouped as follows: i: i u: u e: e, ɛ

	ə: ə, ø a: a, ɑ o: ɔ, o
5	The groups defined in level 4 are further grouped as follows: u: u, o a: a, ə i: i, e

Table 7: Effect of chunking levels on vowels.

Table 8 illustrates the effect of chunking the nasal vowel ɐ̃.

Level	Representation	Comments
1	ɐ̃	Full IPA
2	ɐ	Diacritics dropped
3	ɑ	ɐ merges with ɑ
4	a	ɑ merges with a
5	a	a does not change

Table 8: Illustration of chunking on a single vowel.

The following example illustrates how a piece of Spanish text is represented at each of the five chunking levels.

Original text:

En las primeras dos décadas de existencia de la liga, las competiciones se organizaron básicamente en derredor de las escuelas y clubes de inmigrantes británicos, íntimamente relacionado con las nociones del juego limpio y la caballerosidad deportiva, que constituían el eje de la concepción británica del deporte.

Chunking Level	Transcription
1	en las pɾimeɾas dos dekaðas de egzistenθja de la liɣa las kompetiθjones se oɾɣaniθaɾon basikamente en dereðoɾ de las eskwelas i kluβes de inmiɣɾantes bɾitanikos intimamente ɾelaθjonaðo kon las noθjones del

	xweʏo limpjo i la kaβaʌɾosiðað depoɾtiβa ke konstituian el exe de la konθepθion bɾitanika del depoɾte
2	en las pɾimeɾas dos dekaðas de egzistenθja de la liʏa las kompetiθjones se oɾʏaniθaɾon basikamente en dereðoɾ de las eskwelas i kluβes de inmiʏɾantes bɾitanikos intimamente ɾelaθjonaðo kon las noθjones del xweʏo limpjo i la kaβaʌeɾosiðað depoɾtiβa ke konstituian el exe de la konθepθion bɾitanika del depoɾte
3	en las pɾimeɾas tos tekaθas te eksistenθja te la lixa las kompetiθjones se oɾxaniθaɾon pasikamente en tereθoɾ te las eskʍelas i kluɸes te inmixɾantes pɾitanikos intimamente ɾelaθjonaθo kon las noθjones tel xʍexo limpjo i la kaɸaʌeɾosiθaθ tepoɾtiɸa ke konstituian el exe te la konθepθion pɾitanika tel tepoɾte
4	en las tɾineɾas tos tetasas te etsistensɹa te la lisa las tontetisɹones se oɾsanisaɾon tasitanente en teresoɾ te las estlelas i tluses te innisɾantes tɾitanitos intinanente ɾelasɹonaso ton las nosɹones tel sleso lintɹo i la tasaleɾosisas tetoɾtisa te tonstituian el ese te la tonsetsion tɾitanita tel tetoɾte
5	it sas tritiras tus titasas ti itsistitssa ti sa sisa sas tuttitissutis si ursatisarut tasitatitti it tirisur ti sas istsisas i tsusis ti ittisrattis tritatitus ittitatitti risassutasu tut sas tussutis tis ssisu sittsu i sa tasasirusisas titurtisa ti tutstituiat is isi ti sa tutsitsiut tritatita tis titurti

Table 9: Illustration of chunking a piece of Spanish text.

A similar method was devised by House & Neuberg (1977) who converted the phonemes present in speech data into five very broad phonetic classes. Their method, which they did not ever fully test, then extracted bigram and trigram data from the chunked phonetic information, which would have been used for language identification.

Ringe (1992:67-70) used a very similar method and concluded that "admitting inexact phonological matchings does not make it easier to demonstrate a relationship between languages; at best it should not change the mathematics of the comparison at all." Unfortunately, he based this conclusion on a comparison attempting to show a genetic connection between Navajo and English, which presumably is unlikely to reveal anything, no matter what method is applied.

Kessler (2001:78), in contrast, felt that "there may be situations in which such lumping would give positive results". In carrying out my experiments, I was making no assumption about which chunking levels would provide the greatest accuracy. One of the aims of the first experiments was to determine this objectively.

8.10 Experimental design

Experiments were carried out in a number of phases, each intended to inform the design of the next. The phases were as follows:

1) Simulate evolution, and apply the n-gram comparison method to the resulting artificial languages. Use the n-gram data to generate a possible tree for the languages, and compare with the known model tree.
2) Apply the n-gram comparison method to 10 languages in the Slavic family. Compare the generated tree with the generally accepted Slavic family tree.
3) Apply the n-gram comparison method to a number of native Brazilian languages whose relatedness is not well understood. Use information learned from the first two phases to ensure that the configuration of the system is likely to yield meaningful results.

(This three-part methodology follows Embleton 1986).

Additionally, a number of statistical tests were carried out to ensure that the results were not just a product of chance (as discussed in detail by Ringe 1992) and to determine which factors were significant in producing good results.

8.11 Comparing trees

A major component of my work involved the simulation of language in order to evaluate the effectiveness of the comparison method.

My method involves comparing a number of languages (real or artificial) and using the n-gram distance data to generate a tree (the system automatically generates a tree using the

pair-group method described in section 5.1). In order to automatically determine the accuracy of the trees produced, and to avoid any element of subjectivity, the system automatically determines the generated trees' accuracy, by comparing them to a known model tree. This idea was used by Embleton (1986) and by Ringe, Warnow and Taylor (2002).

I used Embleton's method for determining the accuracy of trees— the topological similarity coefficient (TSC) (Embleton 1986:84). This method involves generating a matrix of distances between pairs of nodes for each of the two trees (the tree generated by lexicostatistics and the model tree) and using these to calculate a single coefficient which represents the degree to which the two trees are topologically similar.

The distance between a given pair of nodes is counted as the maximum number of edges traversed to get from their common ancestor to a terminal node. Each non-terminal node is labeled with its maximum distance, which is used to determine the distance between its common descendants. For example, given the tree in Figure 5 below, the distance between node A and node B is 1 and the distance between A and C is 4. Although the distance between A and the root node (the common ancestor of A and C) is only 2 (i.e., 2 edges) and the distance from C to the root node is 3, the label attached to the root node is 4, which is the distance between E and the root node.

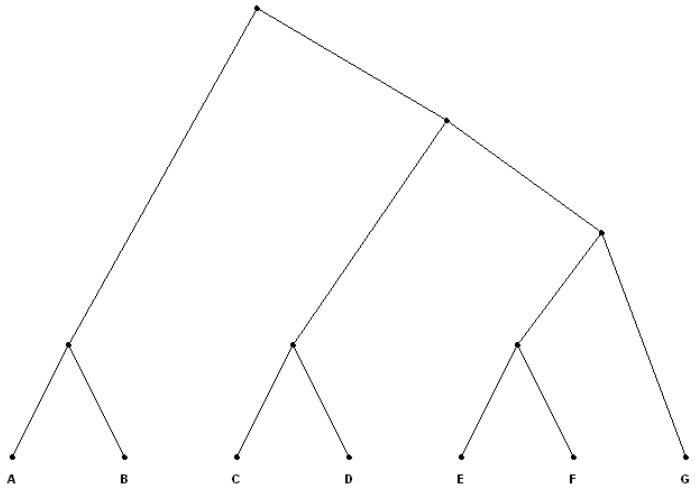

Figure 5: A sample tree.

The distance between each pair of nodes is calculated on this basis, and tabulated in a similarity matrix, as shown in Figure 6.

	A	B	C	D	E	F	G
A	1	**1**	**4**	**4**	**4**	**4**	**4**
B	1	1	**4**	**4**	**4**	**4**	**4**
C	4	4	1	**1**	**3**	**3**	**3**
D	4	4	1	1	**3**	**3**	**3**
E	4	4	3	3	1	**1**	**2**
F	4	4	3	3	1	1	**2**
G	4	4	3	3	2	2	1

Figure 6: The similarity matrix (including redundant values) for the tree in Figure 5.

Note that the distance matrix is symmetrical about its leading diagonal, meaning that the top half above the main diagonal (highlighted in bold) can be stored without including the redundant second half. The leading diagonal is also redundant, as every node is always distance 1 from itself, according to this metric.

To compare two trees the similarity matrix for each is computed, and the matrices are compared using the Pearson Product-Moment Correlation Coefficient (r), defined as follows:

$$r = \frac{1}{n-1}\sum_{i}\left(\frac{X_i - \overline{X}}{s_X}\right)\left(\frac{Y_i - \overline{Y}}{s_Y}\right)$$

n is the number of items in each matrix

$\overline{X}$, $\overline{Y}$ are the mean values in the first and second matrix

X_i , Y_i are the values in the first and second matrix

s_X , s_Y are the sample standard deviation of the values in the first and second matrix

This calculation provides a correlation coefficient which can theoretically range between

-1 and 1 (although Embleton (1986:88) notes that for binary trees the values can in fact only range from -0.5 to 1). A coefficient of 1 indicates that the two trees are topologically identical, while a negative value indicates a total lack of topological correspondence.

8.11.1 Statistical analysis of the TSC

In order to determine what level of TSC score should be considered a close match, some objective statistical analysis was carried out. The first step was to generate 100,000 random trees with the same number of nodes as the test data (i.e. 16 artificially evolved languages, as explained in section 9.1).

Felsenstein (2004:534) points out that it is not a simple matter to decide how to produce random trees for this kind of comparison. In fact, in this case (and in any case where distance data is being used rather than character data), this difficulty does not apply, as it was possible to generate a set of random distance data and use these to produce trees using the pair-group method. Each randomly generated tree was then compared with a model tree (the model tree for the 16 artificially evolved languages—see figure 7 on page 51) and a TSC computed. The distribution of TSC scores generated in this way is shown in Table 10.

TSC range	**count**	**percentage**
-0.5 to -0.4	0	0%
-0.4 to -0.3	0	0%
-0.3 to -0.2	7	0.007%
-0.2 to -0.1	10,700	10.7%
-0.1 to 0.0	45,231	45.2%
0.0 to 0.1	30,909	30.9%
0.1 to 0.2	9,982	10.0%
0.2 to 0.3	2,484	2.5%
0.3 to 0.4	554	0.55%
0.4 to 0.5	109	0.11%
0.5 to 0.6	21	0.02%
0.6 to 0.7	3	0.003%
0.7 to 0.8	0	0%
0.8 to 0.9	0	0%
0.9 to 1.0	0	0%

mean = 0.000435

standard deviation = 0.0906

Table 10: Distribution of TSC scores for trees based on random similarity matrices for 16 languages.

From Table 10 it is clear that a TSC of below 0.3 (accounting for 97.5% of all scores) is not significant. In contrast, a score of above 0.4 can be expected to be obtained by pure chance only once in approximately every 750 trees. Thus, if the system regularly produces trees whose similarity to the model tree, as measured by the TSC, is greater than 0.4, then the method is working well. In fact, scores of greater than 0.3 can only be expected to occur by chance once for every 146 randomly generated trees, so scores of greater than 0.3 can also be considered to be reasonably significant.

9 Results

9.1 Simulated language comparison

For this experiment, Spanish and Armenian were used as the starting languages. The intention was to start with two languages which, although distantly related, are extremely dissimilar. The two initial languages were evolved independently of each other, apart from the effects of borrowing.

In each experiment, 3 generations of descendants were produced, leading from 2 starting languages to sixteen fourth generation descendants. Comparison of these 16 languages was then carried out, a tree generated and this tree was compared to the known model tree. By using simulation rather than working with real languages, the model tree can be known perfectly.

The offspring of a given language were given the same name as their parent, appended with "1" or "2". Hence, the fourth generation languages were named Armenian111, Armenian112, ..., Armenian 221, Armenian222, Spanish111, Spanish 112, ..., Spanish221, Spanish222.

The model tree for experiments 1, 2 and 3 is shown in Figure 7. This is the tree that is explicitly generated by the process of evolution, and so is known to be entirely correct. The method that follows will involve using n-gram comparison to attempt to reconstruct the tree for the 16 languages, and comparing it with the model tree as a measure of the method's accuracy. There are around 6×10^{15} possible rooted binary trees for 16 languages (Felsenstein 2004:24), so the chances of generating this tree by pure chance, or even a tree similar to it, are extremely remote.

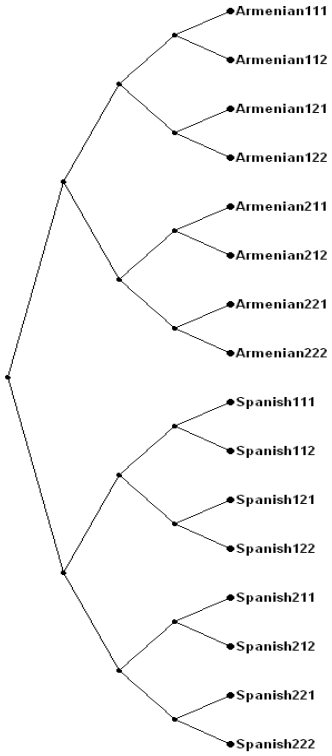

Figure 7: Model tree for 16 languages descended from Spanish and Armenian.

It should be noted that this tree really represents two entirely separate trees. Within this system, there is no common ancestor for Spanish and Armenian. The pair-group method always places all nodes into one tree. Hence, trees generated by this method will tend to indicate that a pair of families or languages are unrelated by having the root node of the tree as their common ancestor.

The main aim of experiments 1-3 was to determine whether the method works and the extent to which borrowing affects its efficacy. Although the method is primarily intended to determine language similarity, these experiments were also testing its ability to determine language relatedness, and thus to sub-group a set of related languages.

9.1.1 Testing the efficacy of the method

First, a simple test was carried out using the settings shown in Table 11.

chunk level	1 (full IPA)
n-gram window	4 (4-grams)
number of iterations of evolution algorithm per generation	100
Borrowing	none
Number of final-generation descendants	16
Volume of text per language	2,000 words

Table 11: Configuration for first test

The tree produced by this test is shown in Figure 8.

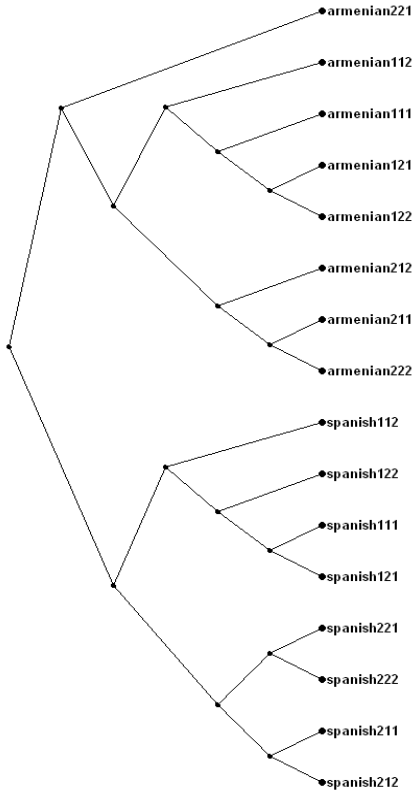

Figure 8: Tree produced by the first simple test.

This tree was automatically compared against the model tree, and obtained a TSC score of 0.91— a very high score, indicating a very close match. Inspection of the tree shows that it is, indeed, a very close match to the model tree. The main split between languages descended from Spanish and those descended from Armenian is correct. The grouping of Spanish221 with Spanish222 and Spanish211 with Spanish211 is perfect. Similarly, all the languages descended from Spanish1 are correctly grouped together. The only error at this level is that Armenian221 is placed as an outlier to the Armenian group, rather than being grouped with the other descendants of Armenian2. Other than this, the split between Armenian1 and Armenian2 is correctly identified.

9.1.2 Borrowing parameters

The next set of experiments were run with 250 iterations per generation, and the following variable borrowing parameters:

- Lexical borrowing allowed (yes / no)
- Phonological borrowing allowed (yes / no)
- Borrowing rate (low / medium / high / very high)

The borrowing rates are defined in Table 12.

Borrowing rate	**Meaning for lexical borrowing**	**Meaning for phonological borrowing**
Low	Replace at most **1%** of words per generation.	Borrow up to **10** phonological innovations per generation.
Medium	Replace at most **5%** of words per generation.	Borrow up to **12** phonological innovations per generation.
High	Replace at most **20%** of words per generation.	Borrow up to **5**0 phonological innovations per generation.
Very high	Replace at most **40%** of words per generation.	Borrow up to **125** phonological innovations per generation.

Table 12: Definitions of borrowing rates.

(It should be noted that in this model a generation is not intended to represent a generation in the traditional sense of human lifetimes. It is simply a single macro-cycle of

the system's algorithm. It is intended to equate to a very long period of time—around 1,000 years, perhaps).

The aim was to determine whether, for example, languages that had evolved without any borrowing could be more accurately sub-grouped by the system than those evolved with a high degree of borrowing.

A total of 11 experiments were run, with the configurations as shown in Table 13.

Experiment number	**Lexical borrowing allowed**	**Phonological borrowing allowed**	**Borrowing rate**	**Restrictions on borrowing**
1i	no	no	N/A	N/A
1ii	yes	no	low	none
1iii	yes	no	medium	none
1iv	yes	no	high	none
1v	no	yes	low	none
1vi	no	yes	medium	none
1vii	no	yes	high	none
1viii	yes	yes	low	none
1ix	yes	yes	medium	none
1x	yes	yes	high	none
1xi	yes	yes	very high	none

Table 13: Configurations for experiments 1i-1xi.

The original text for each language consisted of a 2,000 word sequence of text from www.wikipedia.org. The random number generator used in the system was seeded with a fixed number each time, meaning that although the system had a stochastic basis, each experiment was starting from the same point in the sequence of random numbers. This was an attempt to ensure that the difference between TSC scores was more likely to be

due to real differences in the performance of the system, rather than coincidental random factors.

The results for these experiments are tabulated below. Each experiment was carried out with each combination of chunking levels (1-5) and n-gram window (1-5) providing 25 TSC scores per experiment.

In the tables below, the cells that have scores above 0.4 are shaded. These are scores that are extremely unlikely to occur by pure chance (see section 8.11.1), although scores above 0.30 are also unlikely to occur by chance alone.

Experiment 1i - no borrowing

mean TSC (μ) = 0.62

standard deviation (σ) = 0.16

number of cells that failed to achieve a significant TSC score of 0.40 or above (f) = 3

	1-grams	**2-grams**	**3-grams**	**4-grams**	**5-grams**
chunk =1	0.85	0.66	0.87	0.65	0.64
2	0.41	0.66	0.87	0.65	0.65
3	0.47	0.50	0.65	0.65	0.65
4	0.35	0.74	0.76	0.81	0.74
5	0.38	0.37	0.45	0.50	0.61

Table 14: Results of experiment 1i.

Experiment 1ii - Lexical borrowing, low borrowing rate

$\mu = 0.57$; $\sigma = 0.14$; f = 4

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.65	0.73	0.65	0.59	0.67
2	0.65	0.64	0.65	0.59	0.59
3	0.39	0.45	0.65	0.65	0.67
4	0.31	0.49	0.48	0.74	0.88
5	0.38	0.38	0.41	0.41	0.61

Table 15: Results of experiment 1ii.

Experiment 1iii - Lexical borrowing, medium borrowing rate

$\mu = 0.52$; $\sigma = 0.11$; f = 5

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.52	0.63	0.59	0.55	0.62
2	0.52	0.55	0.59	0.55	0.62
3	0.39	0.41	0.57	0.59	0.65
4	0.35	0.47	0.60	0.57	0.77
5	0.35	0.36	0.32	0.41	0.50

Table 16: Results of experiment 1iii.

Experiment 1iv - Lexical borrowing, high borrowing rate

$\mu = 0.53$; $\sigma = 0.15$; f = 7

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.75	0.61	0.64	0.68	0.68
2	0.72	0.61	0.49	0.61	0.68
3	0.42	0.62	0.61	0.59	0.60
4	0.37	0.60	0.47	0.51	0.37
5	0.25	0.31	0.33	0.36	0.32

Table 17: Results of experiment 1iv.

Experiment 1v - Phonological borrowing, low borrowing rate

$\mu = 0.64$; $\sigma = 0.22$; f = 4

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.52	0.70	0.88	0.83	0.83
2	0.52	0.70	0.77	0.68	0.83
3	0.54	0.88	0.84	0.84	0.88
4	0.57	0.54	0.68	0.79	0.79
5	0.08	0.32	0.30	0.33	0.40

Table 18: Results of experiment 1v.

Experiment 1vi - Phonological borrowing, medium borrowing rate

$\mu = 0.53$; $\sigma = 0.17$; f = 5

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.31	0.71	0.66	0.71	0.85
2	0.31	0.71	0.54	0.55	0.71
3	0.30	0.26	0.42	0.48	0.48
4	0.19	0.55	0.55	0.56	0.66
5	0.40	0.41	0.62	0.62	0.63

Table 19: Results of experiment 1vi.

Experiment 1vii - Phonological borrowing, high borrowing rate

$\mu = 0.51$; $\sigma = 0.24$; f = 9

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.38	0.70	0.69	0.69	0.89
1	0.34	0.57	0.56	0.69	0.89
3	0.25	0.40	0.51	0.66	0.86
4	0.07	0.13	0.43	0.51	0.61
5	0.13	0.36	0.22	0.37	0.80

Table 20: Results of experiment 1vii.

Experiment 1viii - Lexical and phonological borrowing, low borrowing rate

$\mu = 0.48$; $\sigma = 0.26$; f = 9

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.41	0.73	0.73	0.73	0.87
2	0.41	0.64	0.73	0.73	0.82
3	0.36	0.57	0.52	0.52	0.87
4	0.26	0.28	0.25	0.44	0.51
5	-0.05	0.05	0.04	0.29	0.31

Table 21: Results of experiment 1viii.

Experiment 1ix - Lexical and phonological borrowing, medium borrowing rate

$\mu = 0.62$; $\sigma = 0.19$; f = 3

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.46	0.80	0.79	0.79	0.78
2	0.46	0.80	0.79	0.79	0.79
3	0.39	0.45	0.57	0.76	0.79
4	0.24	0.51	0.51	0.51	0.58
5	0.14	0.70	0.69	0.69	0.69

Table 22: Results of experiment 1ix.

Experiment 1x - Lexical and phonological borrowing, high borrowing rate

$\mu = 0.54$; $\sigma = 0.18$; f = 6

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.70	0.77	0.78	0.74	0.70
2	0.70	0.77	0.78	0.74	0.73
3	0.21	0.24	0.47	0.47	0.58
4	0.51	0.43	0.39	0.48	0.42
5	0.46	0.52	0.33	0.33	0.36

Table 23: Results of experiment 1x.

Experiment 1xi - Lexical and phonological borrowing, very high borrowing rate

$\mu = 0.41$; $\sigma = 0.14$; $f = 6$

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.52	0.46	0.42	0.40	0.50
2	0.52	0.46	0.42	0.40	0.50
3	0.54	0.48	0.48	0.49	0.49
4	0.34	0.54	0.50	0.50	0.50
5	0.07	0.27	0.18	0.23	0.10

Table 24: Results of experiment 1xi.

9.1.3 Analysis

A summary of the means and standard deviations of the TSC scores is provided in Table 25.

Experiment number	Lexical borrowing allowed	Phonological borrowing allowed	Borrowing rate	μ	σ	f
1i	no	no	N/A	0.62	0.16	3
1ii	yes	no	low	0.57	0.14	4
1iii	yes	no	medium	0.52	0.11	5
1iv	yes	no	high	0.53	0.15	7
1v	no	yes	low	0.64	0.22	4
1vi	no	yes	medium	0.53	0.17	5
1vii	no	yes	high	0.51	0.24	9
1viii	yes	yes	low	0.48	0.26	9
1ix	yes	yes	medium	0.62	0.19	3
1x	yes	yes	high	0.54	0.18	6
1xi	yes	yes	very high	0.41	0.14	6

Table 25: means and standard deviations of TSC scores for experiments 1i-1xi.

Working on the basis that TSC scores of above 0.4 are highly likely to be significant (i.e. unlikely to have been generated by chance), all of these settings generate results that are significant. Even with very high levels of both phonological and lexical borrowing, the average TSC is 0.41.

A one-way ANOVA (analysis of variance) was used to compare the results of the experiments with high or very high borrowing rates (1iv, 1vii, 1x, 1xi) with the experiment with no borrowing (1i). This test showed a very significant effect at the 1% level ($p<0.01$; $F=67.94$; $df=3$ and 71).

One-tailed paired-sample t-tests were then carried out comparing 1i with the experiments which had high or very high rates of borrowing:

Experiment 1iv - **not significant** at 5% level ($t=1.69$; $df=24$)
Experiment 1vii - **significant** at 0.05% level ($p=0.0005$; $t=8.23$; $df=24$)
Experiment 1x - **significant** at 0.05% level ($p=0.0005$; $t=9.91$; $df=24$)
Experiment 1xi - **significant** at 0.05% level ($p=0.0005$; $t=15.26$; $df=24$)

There is clearly a very high significance in the difference in TSC scores between experiment 1i (no borrowing) and experiments 1vii (high levels of phonological borrowing), 1x (high levels of both phonological and lexical borrowing) and 1xi (very high levels of phonological and lexical borrowing). This suggests (unsurprisingly) that the method performs less well with languages that have evolved with high levels of borrowing than it does with languages that have been reasonably immune to borrowing.

The result of experiment 1iv (high levels of lexical borrowing) is not significantly different from experiment 1i. This suggests that lexical borrowing does not have as great an impact on the accuracy of the system as phonological borrowing. This is not surprising: even with a high level of lexical borrowing, not all words in a language are affected—only those that are replaced—whereas with just a medium rate of phonological borrowing, almost all words in a language can be affected, depending on the selection of

borrowed phonological rules. Additionally, when a language borrows, it is just as likely to borrow from a related language as from an unrelated one. Thus, although a word may be replaced with a new form, this may not affect n-gram scores if the word was already present in the language.

9.1.4 Monte Carlo analysis

The next step was to test whether the results obtained by the n-gram method are really significantly different from those obtained by chance. A Monte Carlo t-test methodology was employed. Repeating t-tests is usually considered to be dangerous, as each repetition increases the likelihood of an error. To mitigate this risk, the following approach was used:

The TSC scores for a given experiment were compared with the randomly selected subset of the 100,000 TSC scores obtained by generating random trees. This comparison was repeated 10,000 times per experiment. Two scores were noted for each experiment: the number of tests for which a significant t-value was obtained and the average t-value. The t-tests were one-tailed, and significance was measured at the 0.05% level. This combination of factors results in an extremely stringent test. (Typically, linguistics papers use a 5% level for significance).

Each t-test compared two independent groups of 25 numbers (5 chunk levels by 5 n-gram windows), one being the fixed set of results for an experiment, and the other a randomly selected subset of the TSC scores obtained for random trees.

Each one-tailed independent samples t-test was carried out 10,000 times, and the t-values compared with the critical value for df=49 (t_{crit}=3.551). The results are shown in Table 26.

Experiment number	Lexical borrowing allowed	Phonological borrowing allowed	Borrowing rate	% significant at p=0.0005	average t (df=49)
1i	no	no	N/A	100%	14.78
1ii	yes	no	low	100%	14.80
1iii	yes	no	medium	100%	14.06
1iv	yes	no	high	100%	12.58
1v	no	yes	low	100%	14.32
1vi	no	yes	medium	100%	11.28
1vii	no	yes	high	99.81%	7.54
1viii	yes	yes	low	100%	15.57
1ix	yes	yes	medium	100%	16.26
1x	yes	yes	high	100%	10.69
1xi	yes	yes	very high	92.9%	5.84

Table 26: Results of Monte Carlo t-test (one-tailed, independent samples) on the TSC scores from experiments 1i-1xi.

The only configurations that performed considerably worse than that without borrowing were 1vii, and 1xi with high levels of phonological borrowing and very high levels of both phonological and lexical borrowing, but all experiments apart from 1xi produced trees that are significantly different from those generated by chance.

It is clear, then that while the system is affected by the level of borrowing, it still produces trees that are significantly more accurate those produced by random distance values even with high levels of borrowing, but when borrowing levels are very high, the trees produced are not significantly different from random trees.

9.1.5 Further analysis of the effects of borrowing rates

In order to provide a more stable analysis of borrowing rates, a further experiment was carried out. This time, chunking level was fixed at 1 (full IPA) and the n-gram window

was fixed at 4 (4-grams). The same configuration settings as experiments 1i-1xi were used, each repeated 5 times, always using the same 10,000 word sections of text. The intention was that by running individual experiments multiple times a truer picture of the variation caused by borrowing could be observed, as opposed to the variation caused by other random factors including selection of text. 100 iterations were used per generation, rather than 250 as was used in experiment 1.

The results, showing five TSC scores per experiment are shown in Table 27.

Experiment number	Lexical borrowing allowed	Phonological borrowing allowed	Borrowing rate	TSC 1	TSC 2	TSC 3	TSC 4	TSC 5	Average
2i	no	no	N/A	0.91	0.85	0.84	0.78	0.79	0.83
2ii	yes	no	low	0.9	0.86	0.84	0.84	0.88	0.86
2iii	yes	no	medium	0.95	0.94	0.87	0.89	0.95	0.92
2iv	yes	no	high	0.66	0.89	0.91	0.82	0.9	0.84
2v	no	yes	low	0.87	0.79	0.85	0.81	0.91	0.85
2vi	no	yes	medium	0.75	0.82	0.85	0.83	0.84	0.82
2vii	no	yes	high	0.83	0.83	0.78	0.89	0.75	0.82
2viii	yes	yes	low	0.9	0.63	0.79	0.76	0.83	0.78
2ix	yes	yes	medium	0.83	0.81	0.83	0.9	0.88	0.85
2x	yes	yes	high	0.69	0.81	0.81	0.51	0.7	0.70
2xi	yes	yes	very high	0.87	0.82	0.49	0.68	0.59	0.69

Table 27: Results of experiments 2i-2xi.

A one-way ANOVA confirmed that the difference between the variances in scores for these 11 experiments was significant at the 0.2% level ($p=0.002$; $F=3.387$; $df = 10$ and 44). A Tukey HSD post-hoc test revealed that the significance was due to the difference between the results of experiment 2x and 2iii, and the difference between 2xi and 2iii. (The same result was obtained using the Bonferroni correction with 95% confidence interval). Interpreting this strictly would suggest that a reasonable amount of lexical borrowing improves the performance of the system, while a large amount of phonological

borrowing decreases it. In fact, since the difference between 2iii and 2i is not significant, it seems more likely that lexical borrowing has had no impact on the quality of the results. It is certainly safe, however, to conclude from these data that phonological borrowing has more impact on the system's accuracy than does lexical borrowing.

9.1.6 Chunking level and n-gram window

The next experiment was designed to determine which chunking level and n-gram window are the best settings to use. The configuration was identical to that used in experiment 1i (i.e., 2,000 words of text, no borrowing, 250 iterations per generation, 16 terminal nodes in the tree). This experiment was repeated 5 times, allowing the random number generator to be seeded randomly each time, leading to a different set of random choices made by the program.

The results were as follows:

Experiment 3i (Repeat of experiment 1i).

mean TSC (μ) = 0.62

standard deviation (σ) = 0.23

number of cells that failed to achieve a significant TSC score of 0.40 or above (f) = 5

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.39	0.47	0.82	0.82	0.83
2	0.39	0.47	0.82	0.83	0.83
3	0.47	0.49	0.83	0.84	0.84
4	0.52	0.73	0.73	0.73	0.80
5	0.07	0.19	0.34	0.42	0.74

Table 28: Results of experiment 3i.

Experiment 3ii (Repeat of experiment 1i).

$\mu = 0.52$; $\sigma = 0.20$; $f = 4$

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.20	0.46	0.60	0.73	0.72
2	0.20	0.46	0.57	0.63	0.72
3	0.40	0.53	0.53	0.73	0.87
4	0.43	0.52	0.58	0.69	0.71
5	0.15	0.15	0.41	0.50	0.63

Table 29: Results of experiment 3ii.

Experiment 3iii (Repeat of experiment 1i).

$\mu = 0.44$; $\sigma = 0.15$; $f = 8$

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.17	0.32	0.46	0.61	0.64
2	0.16	0.14	0.44	0.40	0.42
3	0.35	0.35	0.37	0.44	0.61
4	0.39	0.60	0.59	0.59	0.78
5	0.42	0.41	0.40	0.45	0.45

Table 30: Results of experiment 3iii.

Experiment 3iv (Repeat of experiment 1i).

$\mu = 0.56$; $\sigma = 0.23$; $f = 6$

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.63	0.63	0.69	0.82	0.83
2	0.55	0.63	0.69	0.83	0.83
3	0.40	0.66	0.82	0.63	0.82
4	0.14	0.21	0.21	0.47	0.70
5	0.28	0.28	0.34	0.40	0.40

Table 31: Results of experiment 3iv.

Experiment 3v (Repeat of experiment 1i).

$\mu = 0.42$; $\sigma = 0.17$; f = 12

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.28	0.55	0.55	0.55	0.56
2	0.25	0.30	0.53	0.53	0.59
3	0.38	0.44	0.66	0.69	0.72
4	0.26	0.23	0.12	0.50	0.50
5	0.17	0.26	0.26	0.32	0.34

Table 32: Results of experiment 3v.

It is immediately clear that there is a great deal of variation between the results of experiments 3i-3v, in spite of the fact that they were run with identical configurations. Clearly, then, the evolutionary path followed by a set of languages has an impact on the ability of the n-gram analysis system to correctly analyse the data. This significant variation was confirmed by a one-way ANOVA at the 0.5% level ($p=0.005$; $F=3.985$; df=4 and 120).

A Tukey HSD test showed that the main difference was between 3i and 3v, and between 3i and 3iii. Clearly the languages evolved by experiment 3i were more amenable to analysis than those evolved by 3iii and 3v.

A one-way ANOVA showed a very significant difference between the scores obtained using different chunking levels and n-gram windows, across all 5 experiments ($p<0.0001$; $F=5.043$; df=24 and 100).

The significant differences, according to a Tukey HSD test, were as follows:

- Chunk level 5, 1-grams and 2-grams performed significantly worse than most settings with chunking levels other than 5. (p ranged from 0.000 to 0.050).
- Chunk level 2, 1-grams performed significantly worse than 4-grams and 5-grams with chunk levels of 1 or 3. (p ranged from 0.010 to 0.026).

- Chunk level 1, 1-grams performed significantly worse than 4-grams and 5-grams with chunk level of 1 or 3. (p ranged from 0.023 to 0.047).

This clearly shows that across all the experiments, 1-grams do not perform as well as 4-grams and 5-grams (hardly surprising as 1-grams encode far less information than do 4-grams or 5-grams). It also suggests that chunking level 5 performs less well than other chunking levels. This is presumably due to the loss of information incurred when encoding with chunk level 5.

A look at the average values for each cell across the five experiments is also instructive:

Average for experiments 3i-3v

$\mu = 0.51$; $\sigma = 0.16$; $f = 6$

	1-grams	**2-grams**	**3-grams**	**4-grams**	**5-grams**
chunk =1	0.33	0.49	0.62	0.71	0.72
2	0.31	0.40	0.61	0.64	0.68
3	0.40	0.49	0.64	0.67	0.77
4	0.35	0.46	0.45	0.60	0.70
5	0.22	0.26	0.35	0.42	0.51

Table 33: Average TSC scores for experiments 3i-3v.

It is very clear from Table 33 that 1-grams are not as useful at generating good trees as are the other window settings.

An inspection of the data from experiments 3i to 3v and their average values suggests that 5-grams and 4-grams tend to out-perform 2-grams and 3-grams.

It is harder to pick out the best chunking level, although it seems that levels 1 and 3 provide the best results.

9.2 Initial phoneme comparison

It is not easy to make a direct comparison between my method and other phylogenetic methods, most of which are dependent on cognacy judgments. Ringe's (1992) and Kessler's (2001) methods are the most amenable to direct comparison, although they were not designed with the intention of producing genetic affiliation trees for the languages being compared. They were, rather, designed to answer the question "is language X similar to language Y?" or, to be more precise, "how likely is it that the evidence for relatedness of languages X and Y could have occurred by chance?" This is a useful and important question to answer, but does not help directly with the analysis of large numbers of poorly understood languages.

The method proposed by Ringe (1992) and extended by Kessler (2001) counted the number of recurrences of initial phonemes in cognate forms between languages, and then used statistical tests to determine whether the number of matches identified could have occurred by pure chance.

A modified version of this mechanism can be used to generate trees, and thus can be compared with my n-gram comparison method. This involves treating the percentage of matching initial phonemes as a similarity score between a pair of languages.

This method was incorporated into the simulation framework described above, and the trees it generated were compared with the model tree. The test was run once using each chunking level. For this test, the initial languages used were Spanish and Russian. These languages were chosen simply because comparable Swadesh lists (with 207 items) were available.

The experiment was repeated with different borrowing levels, and the results for the initial phoneme method were compared with the results obtained using the n-gram comparison method. (In the results listed below, cells that scored 0.40 or above are shaded).

Experiment 4i - No borrowing

Chunk Level	TSC
1	0.90
2	0.90
3	0.92
4	0.90
5	0.88

Table 34: Results of experiment 4i using the initial phoneme method.

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.88	0.90	0.89	0.90	0.90
2	0.60	0.73	0.79	0.74	0.90
3	0.40	0.30	0.53	0.53	0.90
4	0.53	0.57	0.70	0.70	0.79
5	0.44	0.46	0.47	0.58	0.58

Table 35: Results of experiment 4i using the n-gram comparison method.

The results of experiment 4i indicate that with no borrowing, the initial phoneme comparison method performs with roughly the same level of accuracy as the n-gram comparison method using a chunking level of 1. A two-tailed independent samples t-test confirmed that the two were not significantly different ($p=0.446$; $n=5$; $df=8$; $t=0.802$).

Experiment 4ii - Medium levels of lexical and phonological borrowing

Chunk Level	TSC
1	0.96
2	0.96
3	0.98
4	0.83
5	0.92

Table 36: Results of experiment 4ii using the initial phoneme method.

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.68	0.81	0.78	0.82	0.84
2	0.43	0.81	0.82	0.84	0.86
3	0.43	0.76	0.83	0.87	0.87
4	0.47	0.69	0.78	0.83	0.83
5	0.36	0.49	0.52	0.61	0.62

Table 37: Results of experiment 4ii using the n-gram comparison method.

The results of experiment 4ii show that with medium levels of borrowing, the initial phoneme comparison method performs significantly better than the n-gram comparison method using 5-grams. This was confirmed using a one-tailed independent samples t-test ($p=0.047$; $t=2.345$; $df=8$).

Experiment 4iii - High levels of lexical and phonological borrowing

Chunk Level	TSC
1	0.69
2	0.69
3	0.67
4	0.83
5	0.81

Table 38: Results of experiment 4iii using the initial phoneme method.

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.62	0.66	0.66	0.66	0.85
2	0.13	0.25	0.32	0.40	0.78
3	0.10	0.17	0.35	0.25	0.50
4	0.16	0.15	0.33	0.41	0.33
5	0.08	0.10	0.18	0.12	0.21

Table 39: Results of experiment 4iii using the n-gram comparison method.

The results of experiment 4iii show that with high levels of borrowing the phoneme comparison method does not perform significantly better or worse than using the n-gram comparison method with chunking level set to 1. This was confirmed using a two-tailed independent samples t-test ($p=0.391$; $t=0.906$; $df=8$).

These results are somewhat disappointing. At best, the n-gram comparison method performs only as well as comparing initial phonemes. At worst, it is significantly out-performed.

9.3 Results of pair-wise n-gram comparison

The results of the pair-wise n-gram comparison method (see section 8.6), applied to the simulated language data used in experiment 4iii (high levels of phonological and lexical borrowing) are shown in Table 40.

Experiment 4iv - High levels of lexical and phonological borrowing, pair-wise n-gram comparison

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.73	0.96	0.96	0.96	0.94
2	0.32	0.81	0.81	0.96	0.58
3	0.34	0.81	0.81	0.96	0.58
4	0.31	0.81	0.81	0.81	0.96
5	0.36	0.34	0.72	0.75	0.96

Table 40: Results of pair-wise n-gram comparison method on simulated language data.

A one-tailed paired samples t-test showed that these results were significantly better than those obtained using the standard n-gram comparison method. ($p<0.0005$; $df=24$; $t=8.01$).

Furthermore, a one-tailed independent samples t-test confirmed that the results for this new method using chunk=1 were significantly better than those produced using the initial phoneme method at the 1% level ($p<0.01$; $t=3.05$; $df=9$).

Hence, although the initial phoneme method performs about as well as the standard n-gram comparison method, the pair-wise n-gram method performs significantly better than either with languages that have been subject to high levels of lexical and phonological borrowing.

9.4 Application to Slavic languages

The next phase of experimentation involved testing the n-gram comparison method with real languages—10 languages from the Slavic family:

West Slavic Languages	South Slavic Languages	East Slavic Languages
Czech	Serbian	Russian
Slovak	Croatian	Ukrainian
Polish	Macedonian	Belarusian
	Bulgarian	

This division is from Sussex & Cubberley (2006:42-54). The model tree for these 10 languages is shown in Figure 9. This tree was derived from a tree in the Encyclopedia Britannica. The decision was taken to group Belarusian and Ukrainian more closely with Russian as a cousin to those two, in line with the trees produced by Nicholls & Gray (2006:168) and Pagel & Meade (2006:176).

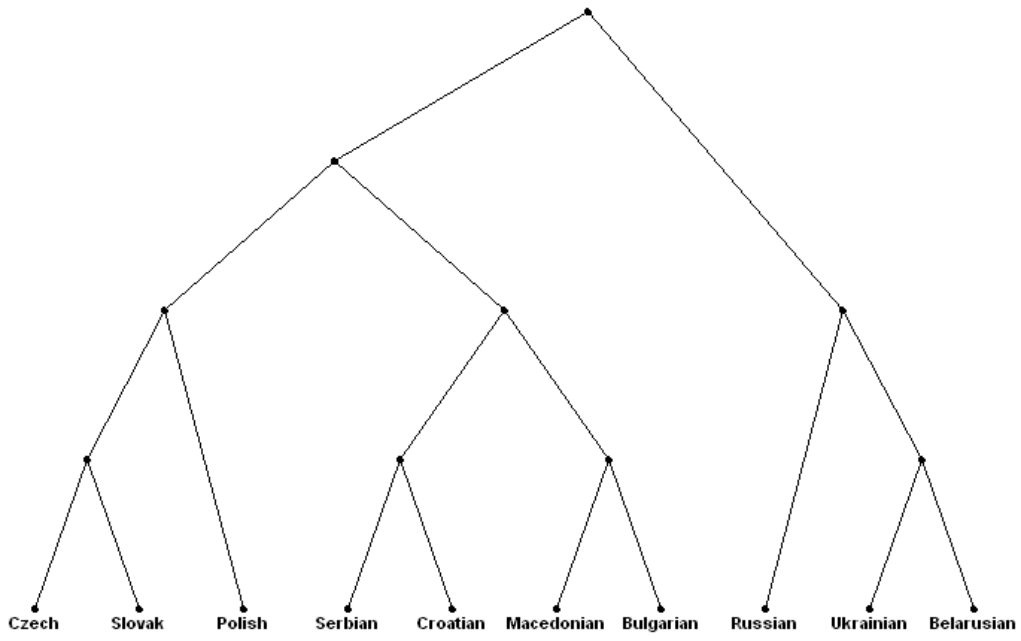

Figure 9: Model tree for the Slavic family

The purpose of these experiments was:

1) To test the system with a set of real languages whose relatedness is reasonably well understood.
2) To measure the effects of varying the length of text used on the effectiveness of the system. Does using a longer piece of text provide better results?

The Slavic family was chosen for the following reasons:

1) The Slavic languages mainly have fairly shallow orthographies (see section 8.8), meaning that it was not difficult to develop rules for converting their texts into IPA notation.
2) According to Paulston and Peckham (1998:258): "All the Slavic languages are mutually intelligible to a degree". Although most of the Slavic languages are relatively similar to each other, there are some pairs (particularly Serbian / Croatian, Macedonian / Bulgarian and Czech / Slovak) which are considered by some to be dialects of the same language. This provides a test of the system's ability to work with languages that are very similar to each other. This is, for example, a much harder task than sub-grouping the following set of languages: Spanish, Italian, Serbian, Croatian, Greek, Cypriot Greek.

Since these experiments were run with a different set of languages from those in the simulations, a different model tree was used. Hence, it was important to recalculate the table of TSC scores for randomly generated trees. This is because a randomly generated tree with 10 terminal nodes has a different chance of matching a given tree from one with 16 nodes.

100,000 sets of randomly generated distance data were produced, and each was used to produce a tree using the pair-group method described on page 17. Each tree was compared with the Slavic model tree and a TSC score generated (see section 8.11). The distribution of these scores is shown in Table 41.

-0.5 to -0.4	0	0%
-0.4 to -0.3	0	0%
-0.3 to -0.2	4,558	4.6%
-0.2 to -0.1	23,594	23.6%
-0.1 to 0.0	28,453	28.5%
0.0 to 0.1	21,147	21.1%
0.1 to 0.2	12,430	12.4%
0.2 to 0.3	5,314	5.3%
0.3 to 0.4	2,742	2.7%
0.4 to 0.5	1,217	1.2%
0.5 to 0.6	307	0.3%
0.6 to 0.7	133	0.1%
0.7 to 0.8	88	0.09%
0.8 to 0.9	14	0.01%
0.9 to 1.0	3	0.003%

mean = 0.000077

standard deviation = 0.15

Table 41: Distribution of TSC scores for trees based on randomly generated similarity matrices for 10 Slavic languages.

As with the 16 language data, a TSC of below 0.3 is not significant, accounting for 95.5% of the generated trees. A score of 0.5 can be expected to be obtained by chance once in approximately every 200 trees. Thus, if the system regularly produces trees whose similarity to the Slavic model tree, as measured by the TSC, is greater than 0.5, then the method is working well. Scores of over 0.4 can also be considered to be reasonably significant.

The main factors varied in this set of experiments were the chunking level, n-gram window and length of text.

The first four experiments were identical, but using different lengths of texts, randomly selected from a sequence of text from Wikipedia. For experiment 6, Swadesh lists of 207 words were used (see Appendix A - The Swadesh lists). Although these lists are very

short, they provide an interesting cross-over with the standard lexicostatistical technique, as they rely entirely on data obtained from relatively basic vocabulary and are thus likely to provide a certain degree of immunity to effects of borrowing.

Results were as follows (cells in the tables with scores of 0.50 or above are shaded. Cells with a score between 0.40 and 0.50 are given a lighter shading. Unshaded cells represent insignificant results).

Experiment 5i - 250 words

$\mu = 0.24$ (excluding 5-grams)

$\sigma = 0.21$ (excluding 5-grams)

f = 19 (below 0.40 - excluding 5-grams)

	1-grams	**2-grams**	**3-grams**	**4-grams**	**5-grams**
chunk =1	0.32	0.31	0.25	0.19	N/A
2	-0.12	0.14	-0.08	0.10	N/A
3	-0.05	0.28	0.26	0.27	N/A
4	0.33	0.62	0.05	0.23	N/A
5	0.05	0.27	-0.07	0.17	N/A

Table 42: Results of experiment 5i.

The method was unable to generate a tree for 5-grams using 250 word texts. This was because in each case, every pair of languages was deemed to have a similarity score of 1 (i.e., no similarity at all). A tree can be generated from this data, but it is the same as producing a tree from a set of random scores.

It is clear by inspection that these results are no better than those obtained by chance. There is only one configuration that has a score above 0.4, which is slightly better than would be expected by chance, but hardly significantly so. Clearly, working with 250 random words of text is not adequate for this method. Furthermore, using such a small number of words means that the text that happens to be chosen has a large influence on

the results. For example, selecting a different set of 250 words (from the same corpus) and with all other parameters identical produced the following results:

$\mu = 0.37$ (excluding the N/A values)
$\sigma = 0.25$ (excluding the N/A values)
$f = 13$ (excluding the N/A values)

	1-grams	**2-grams**	**3-grams**	**4-grams**	**5-grams**
chunk =1	0.39	0.63	0.70	0.98	N/A
2	0.11	0.63	0.36	0.41	N/A
3	-0.05	0.20	0.09	0.18	N/A
4	0.38	0.53	0.54	0.38	0.11
5	0.00	0.17	0.47	0.38	0.46

Table 43: Results of experiment 5i, using a different set of data.

Although, on average, these results are still not significantly better than those produced by random trees ($\mu = 0.37$), they do clearly contain some very accurate trees. In particular, the tree produced with chunk level 1 and 4-grams has a TSC of 0.98—the highest obtained for any tree in this study. The tree produced differed from the model tree only in the grouping of Belarusian, Ukrainian and Russian, as shown in Figure 10.

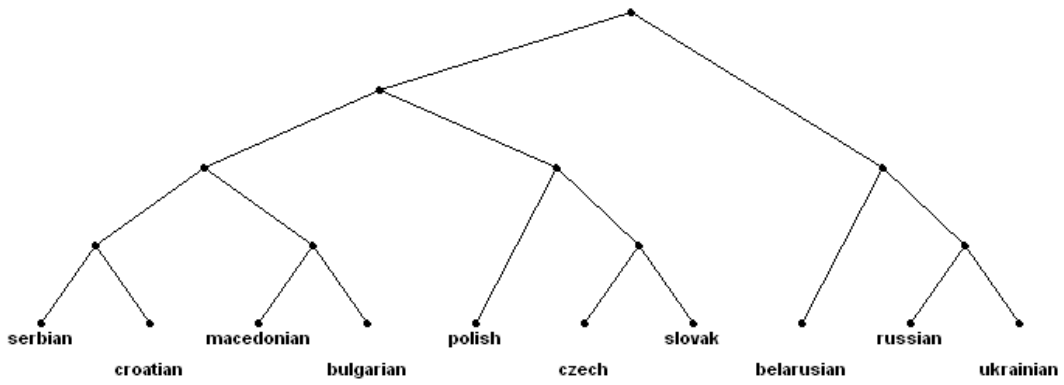

Figure 10: Tree produced using chunk=1 and window=4; TSC=0.98.

It is unsurprising that such variability can result with such small amounts of text. A 4-gram comparison, for example, of two 250 word texts is unlikely to yield the same result when repeated with different texts from the same languages. As the volume of text increases, so the n-gram vector for the text becomes more representative of the language itself rather than just a specific text.

Since it is unlikely to be possible to select the "magic" text which produces the correct tree, the best way to apply the n-gram comparison method is with larger amounts of text. This can be seen from the results of experiments 5ii, 5iii and 5iv:

Experiment 5ii - 2,000 characters

$\mu = 0.24$; $\sigma = 0.21$; f = 19 (below 0.40)

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.08	0.39	0.23	0.41	0.64
2	0.08	0.02	0.12	0.11	0.34
3	0.02	0.03	0.02	0.19	0.65
4	0.14	0.42	0.59	0.16	0.62
5	0.03	0.12	0.15	0.38	0.18

Table 44: Results of experiment 5ii.

Experiment 5iii - 10,000 characters

$\mu = 0.25$; $\sigma = 0.23$; f = 19

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.08	0.23	0.26	0.41	0.33
2	0.03	0.08	0.10	0.10	0.29
3	0.03	0.05	0.03	0.23	0.00
4	0.26	0.24	0.43	0.86	0.85
5	0.15	0.48	0.48	0.19	0.07

Table 45: Results of experiment 5iii.

Experiment 5iv - 10,000 words

$\mu = 0.30$; $\sigma = 0.20$; f = 14

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.08	0.23	0.47	0.41	0.50
2	0.08	0.07	0.08	0.08	0.24
3	0.03	0.04	0.03	0.23	0.34
4	0.52	0.56	0.40	0.57	0.57
5	0.17	0.49	0.43	0.39	0.50

Table 46: Results of experiment 5iv.

Experiment 6 - Swadesh lists

$\mu = 0.20$; $\sigma = 0.14$; f = 20

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.48	0.39	0.40	0.40	0.40
2	0.03	0.03	0.04	0.11	0.25
3	0.00	0.11	0.04	0.04	0.39
4	0.18	0.20	0.20	0.20	0.18
5	0.18	0.22	0.21	0.2	0.22

Table 47: Results of experiment 6.

Experiment 7 - Pair-wise n-gram comparison method with Swadesh lists

$\mu = 0.50$; $\sigma = 0.16$; f = 9

	1-grams	2-grams	3-grams	4-grams	5-grams
chunk =1	0.73	0.81	0.45	0.45	0.45
2	0.32	0.32	0.38	0.53	0.53
3	0.30	0.30	0.30	0.30	0.30
4	0.61	0.67	0.61	0.61	0.46
5	0.37	0.67	0.66	0.66	0.66

Table 48: Results of experiment 7.

The results of experiment 7 confirm that the pair-wise n-gram comparison method is more effective than the standard n-gram comparison approach. Indeed, at first glance, the results of experiments 5 and 6 do not look promising. However, they are not, in fact, as bad as they appear. An inspection of some of the trees generated by these experiments is instructive. The tree shown in Figure 11 is that produced by experiment 5iv (10,000 words) with chunk=1 (full IPA) and n-gram window = 4 (i.e. 4 grams).

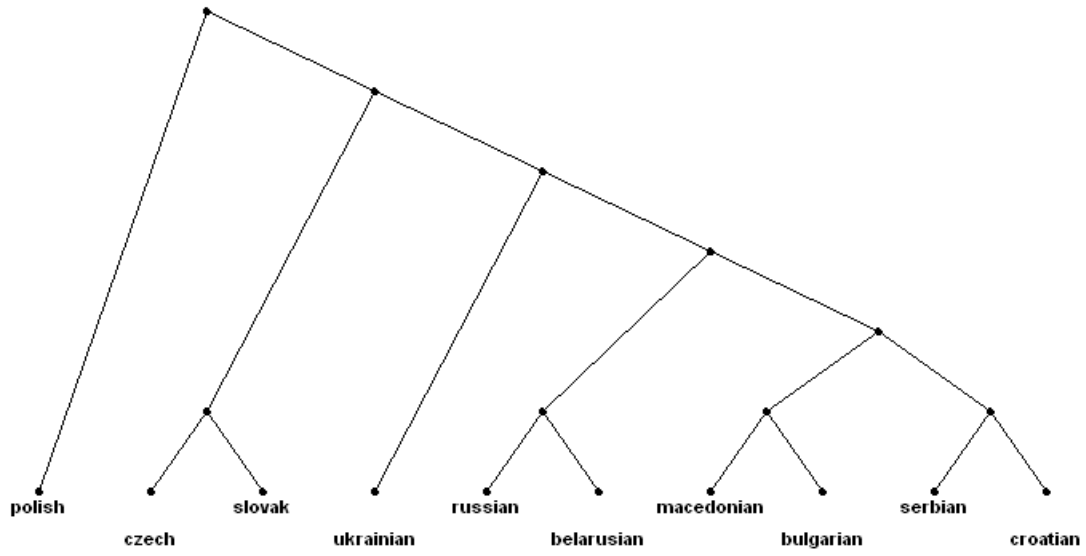

Figure 11: Tree produced by experiment 5iv with chunk=1, window=4; TSC=0.41.

This tree scored a TSC of 0.41, which is only just in the range being considered significant for this experiment. However, the tree is clearly fairly accurate. It correctly identifies the West Slavic / South Slavic / East Slavic divide. The only real differences between this tree and the model tree are:

- It groups Russian and Belarusian more closely than Ukrainian and Belarusian. It is not at all clear which of these is correct, and as is explained above, the model tree's version of these three languages is somewhat arbitrary.
- It places Ukrainian as an outlier to the East Slavic group.
- Although it groups Polish with the other West Slavic languages, it places it as an

outlier to the main tree.

- It groups South Slavic more closely with East Slavic, rather than with West Slavic.

It is thus far from clear that the tree could be described as being materially incorrect; it simply differs from the model tree more than is expected. It seems clear, then, that while the TSC is an objective measure of the extent to which two trees are similar, it does not always map well onto subjective perceptions.

A Monte Carlo independent-samples t-test (one-tailed) was carried out to test the hypothesis that the trees generated by the system for the Slavic languages were significantly better than those obtained by chance. As in section 9.1, the results were compared against 10,000 randomly selected sub-set of the 100,000 TSC scores obtained for random trees.

Experiment	**Size of text**	**% significant at p=0.0005**	**average t value (df = 49)**
5ii	2,000 characters	93.77%	4.78
5iii	10,000 characters	92.27%	4.56
5iv	10,000 words	99.82%	6.09
6	207 words (Swadesh list)	91.96%	5.08
7	207 words (Swadesh list)	100%	11.45

Table 49: Results of Monte Carlo t-test (one-tailed, independent samples) of TSC scores from experiments 5, 6 and 7.

The results in Table 49 show clearly that the trees generated using 10,000 words of data are significantly better than those obtained by pure chance with a probability of 99.82%. Further, the results generated using the pair-wise n-gram comparison method (experiment 7) are significantly better than those produced by the standard n-gram comparison method.

9.4.1 Diacritics

The final conclusion to be drawn from this set of experiments is that using a chunking level of 1 (in other words, using the full IPA encoding) with 4-grams is the most effective configuration for the Slavic languages. This produced a tree with a score of 0.40 or above in all of experiments 5, 6 and 7 (with the exception of 5i which used only 250 words of text per language). It is surprising, at first glance, that using chunking level 2 performs so much worse than chunking level 1.

The diacritics used in the Slavic languages included in these experiments are shown in Table 50.

Diacritic	Languages	Example
palatalisation	Polish, Ukrainian, Russian, Belarusian	tʲ
nasalisation	Polish	ã
raised	Czech	r̝
long	Russian, Czech, Slovak	ɕː

Table 50: Diacritics used in the phonemes of the 10 Slavic languages used in experiments 5, 6 and 7.

The difference between chunking level 1 and chunking level 2, then, is entirely in these four diacritics. Since nasalisation and the raised trill are only present in Polish and Czech respectively, it seems unlikely that these are involved. Indeed, an examination of the trees produced by the system suggest that most of the difference between trees produced with chunking level 1 and those produced with chunking level 2 are due to Russian. Specifically, moving from level 1 to level 2 moves Russian closer to Serbian and Croatian, and further from Belarusian. It is likely, then, that this reflects the fact that there

is not a great deal of variation between the 10 Slavic languages included in this study and that some of the differences that do exist are not good indicators of genetic affiliation.

These findings go some way to confirming Kessler's (2001:79) prediction that treating palatalized consonants as their plain equivalents would produce better results for some languages and worse for others.

9.5 Application to native Brazilian languages

A set of 100 word Swadesh lists was obtained from the web site of the Grupo de Investigação Científica de Línguas Indígenas (GICLI - http://paginas.terra.com.br/educacao/GICLI/ListasEnglish.htm - accessed between 21st and 26th May 2008) for 29 Brazilian aboriginal languages. The languages used are listed in Table 51.

1. Aikanã
2. Akawaio
3. Amondava
4. Apalaí
5. Apinayé
6. Apurinã
7. Arára Pano
8. Arikapú
9. Ashéninka
10. Asuriní Do Tocantins
11. Asuriní Do Xingu
12. Awetí
13. Bakairí
14. Baniwa
15. Borôro
16. Cinta-Larga
17. Dâw
18. Dení
19. Guajajára
20. Guarani Antigo
21. Guató
22. Hixkaryána
23. Ikpeng
24. Pirahã
25. Urubú-Kaapor
26. Yaminawa
27. Yanam
28. Yawalapiti
29. Yawanawa

Table 51: List of native Brazilian languages.

The purpose of this test was to demonstrate the real value of my method: in particular, that it does not rely on the comparative method, and that it can be automated with very little effort. In this case, all that was available for each language was an IPA-encoded list of up to 100 words per language. Given this limited set of data it would be impossible to apply most traditional lexicostatistical methods, and even methods such as Heggarty's inspection method could not be applied. Ringe's and Kessler's inspection methods based primarily on initial characters could be applied, in theory (but see limitations discussed below), but would not provide as rich an analysis of the data as n-gram analysis.

The system was run on the texts of the 29 languages with chunking set to 1 (i.e. making full use of the IPA characters) and n-gram window set to 4 (4-grams). These settings were chosen on the basis of evidence gathered during the earlier experiments.

The results were examined by hand. It was not feasible to generate a TSC score since no accepted model tree exists for these languages. The tree generated by the system is shown in Figure 12. There are nearly $9x10^{36}$ possible trees for 29 languages (Felsenstein 2004:24), so the chances that this tree even approximately matches the "correct" tree (if such a tree could be said to exist) by pure chance are non-existent.

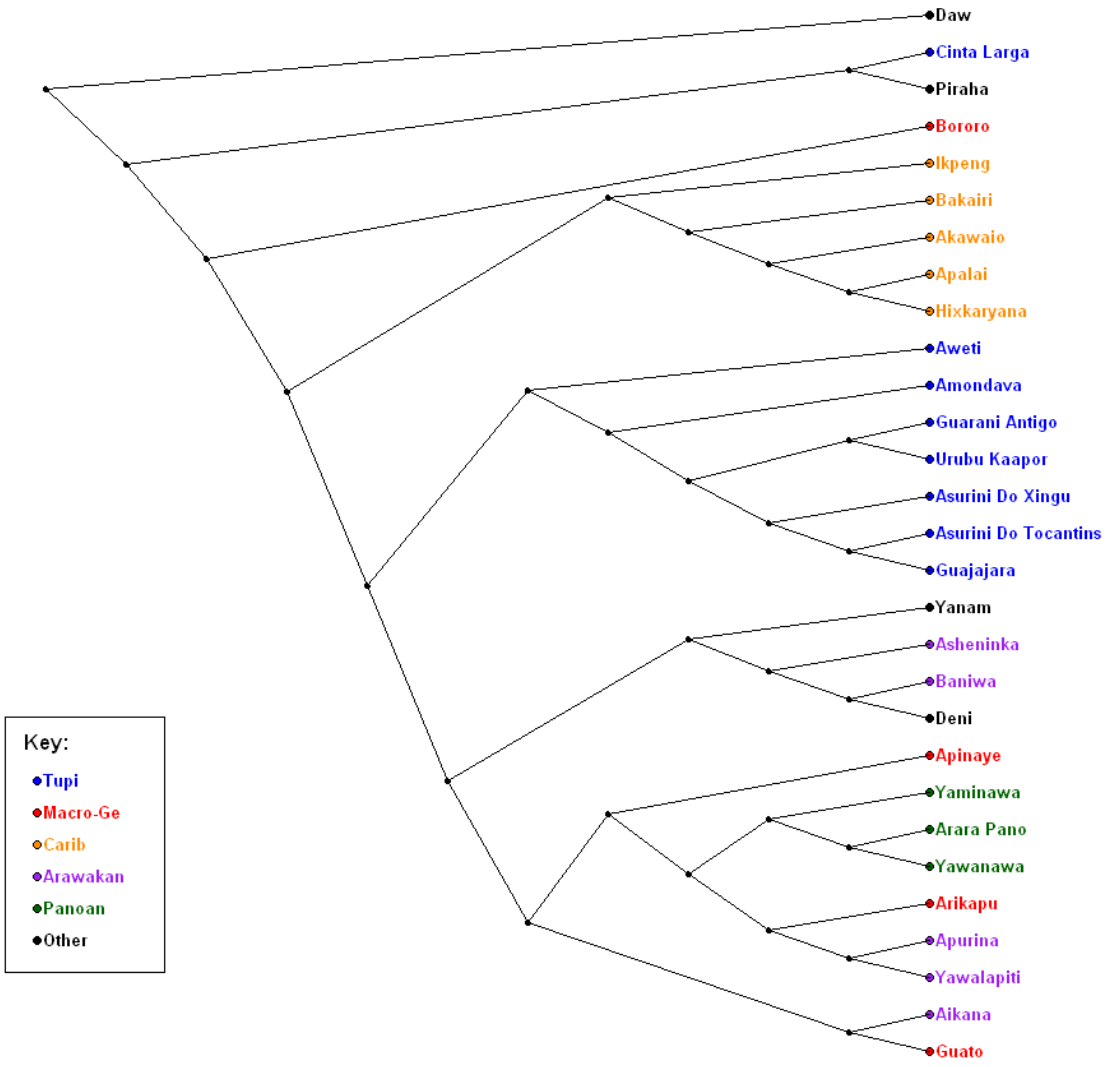

Figure 12: Tree generated for 29 Brazilian Aboriginal languages using 4-grams, with chunk=1.

9.5.1 Pirahã and Cinta-Larga

In Figure 12 Pirahã is sub-grouped with Cinta-Larga. According to Gordon (2005), Cinta-Larga is related to the Mondé sub-group of the Tupi languages. There is no reason to believe that Pirahã is related to this group, but it is instructive to examine what further information can be obtained from the present system to see whether a strong case can be made for a relationship.

Using an n-gram window of 5 (with chunking set to 1) Table 52 shows a complete list of the matching 5-grams between Cinta-Larga and Pirahã:

5-gram	**Matching meanings in Pirahã list**	**Matching meanings in Cinta-Larga list**
i̯kop	black	know, say
i̯ʔaa	root, foot, bite, dry	lie down
i̯kopa	black	know
ii̯ka	water, rain, ash, cold, big, woman, tree, flesh, moon, who, mouth, name, full	hear, die, swim, bone, round

Table 52: All matching 5-grams between Cinta-Larga and Pirahã.

It is fairly clear from this list that the 5-gram matches are extremely likely to be coincidental. A similar picture emerges from the matching 4-grams.

This test thus does not provide any evidence of a relationship between Pirahã and Cinta-Larga, although a more detailed investigation may be warranted; the evidence provided here does not in any way prove that the languages are *not* related, it simply does not provide any convincing evidence that they are. The important point is that my method provided a short-cut to determine which pairs of languages might be worth examining further. In this case, the fact that Pirahã is not known to have any non-extinct related languages means that any suggestion of a relationship is of interest, and worth examining further.

A similar comparison carried out for another pair of languages apparently closely related (Asuriní do Tocantins and Guajajára) provides a much larger list of matching 5-grams, a selection of which are shown in Table 53.

5-gram	**Matching meanings in Asuriní do Tocantins list**	**Matching meanings in Guajajára list**
wɨtɨr	mountain	mountain
ˌpɨʔa	liver	liver
ahɨta	star	star
ŋˌnam	ear	ear
ˌmoko	two	two

Table 53: Examples of matching 5-grams between Asuriní do Tocantins and Guajajára.

It is clear from this list that there is a strong case for a genetic connection between these languages, particularly given the basic nature of some of the words that are matched. Indeed, an inspection of the word lists shows many words that are identical in form, and many that show regular correspondences such as those shown in Table 54.

English	**Asuriní do Tocantins**	**Guajajára**
big	oho	uhu
long	poko	puku
child	konomi	kulumi
to stand up	puʔen	poʔom

Table 54: Examples of regular correspondences between Asuriní do Tocantins and Guajajára.

Indeed, Guajajára and Asuriní do Tocantins are very closely related; according to Gordon (2005) and Campbell (1997:200) they are both members of sub-group of the Tupi-Guarani family.

9.5.2 Sub-groupings

According to information from Gordon (2005), most of the sub-groupings contained in the tree in Figure 12 are accurate. Working from the top of the tree, the first four languages are clearly separated from the others. This is reasonable for Dâw (a Maku language) and Pirahã (the only surviving Mura language), but Cinta-Larga and Borôro should probably be grouped with the Tupi and Macro-Ge families, respectively, although Campbell (1997:326) says of Macro-Ge that the evidence Greenberg and others have presented so far do not support the Macro-Ge grouping.

The next group, from Ikpeng to Hixkaryána is the Carib sub-group, accurately identified by the system. Next is the Tupi family, including the languages from Awetí to Guajajára. Not only are these 7 languages correctly grouped together, but the final two, forming a close pair, Asuriní do Tocantins and Guajajára are indeed very closely related, from the same sub-group within the Tupi family (the Tenetehara group; Campbell 1997:200).

Yanam, a Yanomaman language, apparently unrelated to any of the other languages, is an outlier to the next sub-group. Next are the Arawakan and Panoan families, slightly mixed together, along with three of the Macro-Ge languages. Dení is a member of the Arauan family, believed to be related to Arawakan (Campbell 1997:178).

The result is remarkably accurate. More than half of the languages are correctly grouped with languages to which they are related, and four major groups are very neatly picked out. Dâw, not related to any of the other languages being examined, is correctly placed as an outlier.

9.5.3 Similarity scores

In fact, the n-gram comparison method was not designed for sub-grouping in the way that it has been used here, as it is really a measure of language similarity, rather than language relatedness. Clearly in this case the two are close enough that the method has worked well. It is now instructive to look more closely at the language similarity scores generated for this test.

The 10 pairs of languages deemed most similar to each other by their n-gram similarity scores are shown in Table 55.

Language pair	**Language Family**	**4-gram similarity score (0 = identical)**
Arára Pano & Yawanawa	Panoan	0.58
Arára Pano & Yaminawa	Panoan	0.60
Yaminawa & Yawanawa	Panoan	0.66
Asuriní do Tocantins & Guajajára	Tupi	0.75
Guarani Antigo & Urubú-Kaapor	Tupi	0.79
Guajajára & Urubú-Kaapor	Tupi	0.79
Asuriní do Tocantins & Asuriní do Xingu	Tupi	0.80
Asuriní do Xingu & Urubú-Kaapor	Tupi	0.81
Asuriní do Xingu & Guajajára	Tupi	0.81
Apurinã & Yawalapiti	Arawakan	0.82

Table 55: The ten most similar language pairs.

Each of these is indeed a related pair of languages, indicating that the method has performed extremely well at identifying related languages.

Table 56 shows, for each language in the test, the other language deemed most similar to it, in order of n-gram similarity scores. From this data, it can be observed immediately that the putative relationship between Pirahã and Cinta-Larga is in fact merely a consequence of the fact that Pirahã is not similar to any of the languages. Like Daw, another language that is not related to any other in the list, its closest neighbour is extremely dissimilar (a score of 1 means no similarity at all). Pirahã and Cinta-Larga score 0.96, and Daw scores 0.97 with its nearest neighbour.

Language	**Closest neighbour**	**Families**	**4-gram similarity score (0 = identical)**
Yawanawa	Arára Pano	Panoan	0.58
Arára Pano	Yawanawa	Panoan	0.58
Yaminawa	Arára Pano	Panoan	0.60
Asuriní Do Tocantins	Guajajára	Tupi	0.75
Guajajára	Asuriní Do Tocantins	Tupi	0.75
Guarani Antigo	Urubú Kaapor	Tupi	0.79
Urubú Kaapor	Guarani Antigo	Tupi	0.79
Asuriní Do Xingu	Asuriní Do Tocantins	Tupi	0.81
Apurinã	Yawalapiti	Arawakan	0.82
Yawalapiti	Apurinã	Arawakan	0.82
Amondava	Guajajára	Tupi	0.85
Awetí	Guajajára	Tupi	0.86
Hixkaryána	Apalaí	Carib	0.86
Apalaí	Hixkaryána	Carib	0.86
Akawaio	Apalaí	Carib	0.86
Arikapú	Yawalapiti	Macro-Ge & Arawakan	0.86
Baniwa	Yawalapiti	Arawakan	0.87
Ashéninka	Apurinã	Arawakan	0.87
Dení	Baniwa	Arauan & Arawakan	0.88
Aikanã	Baniwa	Arawakan	0.90
Apinayé	Arikapú	Macro-Ge	0.90
Bakairí	Hixkaryána	Carib	0.90
Yanam	Yawanawa	Yanomaman, Panoan	0.91
Ikpeng	Yawanawa	Carib, Panoan	0.91
Cinta Larga	Apurinã	Tupi, Arawakan	0.92
Guató	Yawalapiti	Macro-Ge, Arawakan	0.92
Borôro	Guató	Macro-Ge	0.95
Pirahã	Cinta Larga	Mura, Tupi	0.96
Daw	Awetí	Maku, Tupi	0.97

Table 56: List showing the closest pairing for each language.

In table 56, rows where the two languages are not in the same family are shaded, although it should be noted that Dení and Baniwa are well paired, as their families (Arauan and Arawakan) are considered to be related. (Campbell 1997:182). Excluding this pairing, only 7 of the 29 languages are most closely paired with a language that is not from the same family. Of these, three are from families that are not represented elsewhere in the list, so do not represent errors (by its very nature the system must pair each language with another, and if a language is an isolate, or not related to any others in the list, it must be paired with an unrelated language). Additionally, Campbell (1997:204) reports that the Yanomaman family is believed by some to be related to the Panoan group. The only pairings, then, that appear to be genuine errors are shown in Table 57.

Arikapú	Yawalapiti	Macro-Ge & Arawakan	0.86
Ikpeng	Yawanawa	Carib, Panoan	0.91
Cinta Larga	Apurinã	Tupi, Arawakan	0.92
Guató	Yawalapiti	Macro-Ge, Arawakan	0.92

Table 57: The four pairings that appear to be genuine errors.

The first of these pairs, Arikapú and Yawalapiti, is the only incorrect pair to score below (better than) 0.90. An examination of the matching 4-grams between these two languages reveals some potentially interesting similarities, as shown in Table 58.

Matching n-grams	**Yawalapiti**	**Arikapú**
tʃit	tʃitʃu - belly wittʃitʃu - star	tʃitʃi - big
tuka	tuka - drink	tuka - fat / grease
ʃiu‿	puʃiu - bird	patʃiu - night
‿tʃa	tʃa - eat	tʃako - mouth
‿kam	kama - to die	kamu - new
a‿kɨ	kɨtsɨkɨ - sand	kɨkɨra - sand

Table 58: Matching 4-grams between Yawalapiti and Arikapú.

These data are not enough to provide any strong evidence for relatedness, but they are certainly sufficient to warrant further investigation to see if any of the matches identified

here form part of a regular correspondence. One such possible correspondence, suggested by the strong similarity between the words for "sand" is shown in Table 59.

English	**Yawalapiti**	**Arikapú**
sand	kɨtsɨkɨ	kɨk̲ɨra
mouth	kaɲatsi	tʃak̲o
tooth	tsɨwɨ	tʃuk̲rihã
heart	kaɲɨtsi	mək̲a
to walk	tsuka	k̲ərəj
new	autsa	k̲ama

Table 59: Possible regular correspondences between Yawalapiti and Arikapú.

Again, these data are not enough to be compelling evidence, but they are enough to warrant further investigation.

It is worth noting that the genetic classifications assumed here are not at all uncontroversial. Campbell (1997:204) lists a number of proposed remote genetic relationships, including:

1) Carib & Arawakan
2) Carib & Tupi
3) Carib, Tupi & Arawakan
4) Macro-Ge, Pano & Carib

Hence, the fact that the four Macro-Ge languages in this study are grouped with Arawakan, Panoan and Tupi languages may not be entirely unreasonable.

It seems clear, then, that the present method, which took a few hours to set up and less than a minute to run, is capable of producing a tree of genetic affiliation that is extremely close to the accepted view and may well be capable of contributing to the ongoing work of understanding the relatedness of native South American languages by proposing new avenues for investigation. The results should not be considered to have the authority of a detailed investigation using the comparative method, but given the ease with which the

test was run and the fact that it can be automated for large numbers of languages, it clearly has the potential to be an extremely valuable tool to complement more traditional methods, in contrast with other lexicostatistical techniques which can usually only be applied *after* the traditional methods have been applied. This section has also shown that it can work surprisingly well with very meagre data (fewer than 100 words per language - albeit carefully chosen words).

9.5.4 Comparison with the initial phoneme method

Using the initial phoneme comparison method on the Brazilian data was problematic. Although the lists contain up to 100 words, there are only 54 words which are present in the list of every language. This reduced the effective text from 100 words per language to 54. The n-gram comparison method does not require the words to be in a particular order (although it almost certainly is advantageous if they are in roughly the same order in each list) and is not adversely affected by missing words.

In spite of these difficulties, the initial phoneme comparison method was tested with the Brazilian data. For each pair of languages the number of words that had the same initial phoneme was counted. This number was used to compute a percentage similarity score between each pair of languages, and a tree generated as usual. The tree produced by this method is shown in Figure 13.

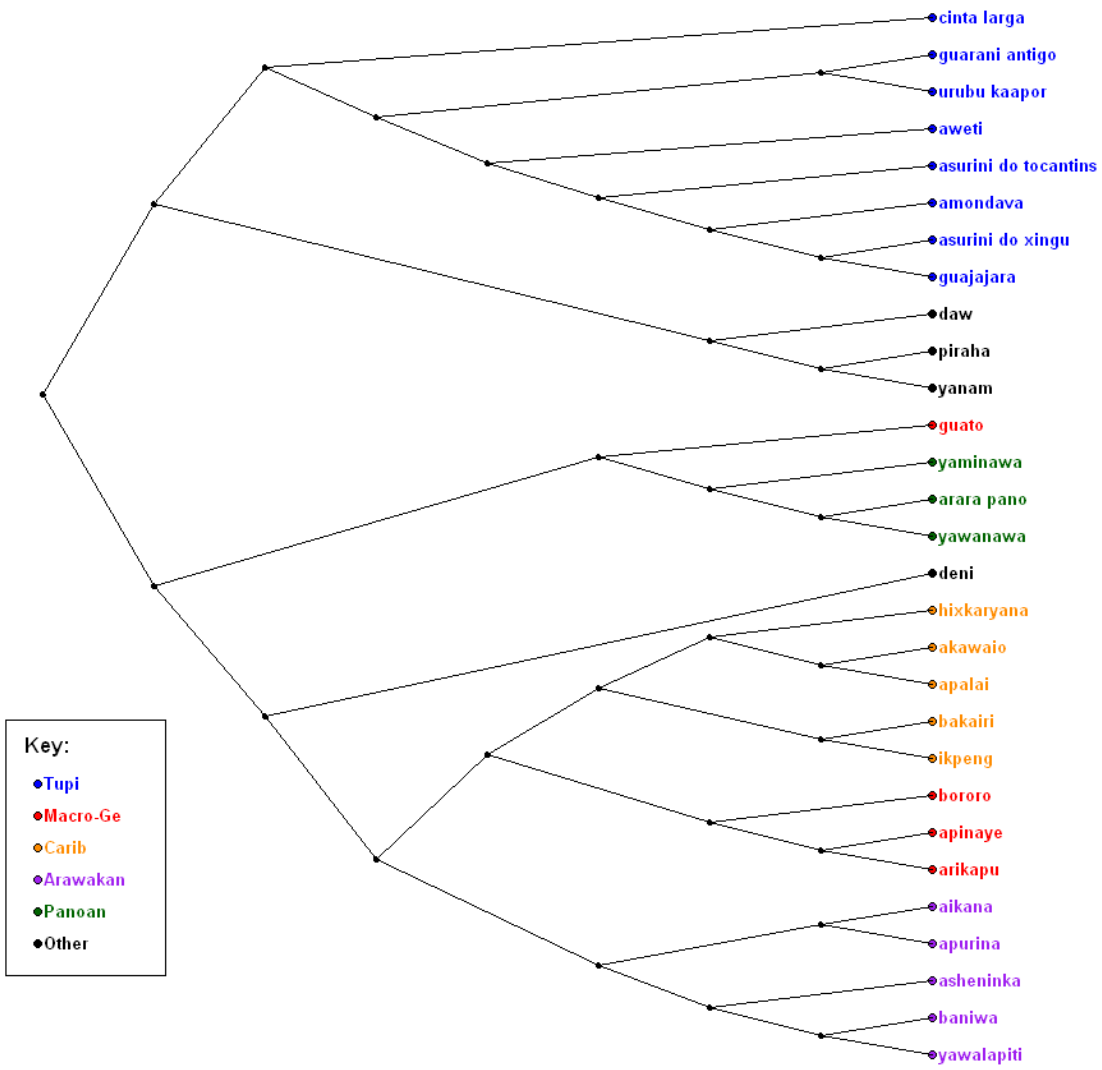

Figure 13: Tree produced by applying initial phoneme comparison to the Brazilian data.

This tree is at least as accurate (compared with the generally accepted view) as the tree generated using the n-gram comparison method. In fact, it improves on that method by grouping Cinta Larga with the other Tupi languages, and by grouping Borôro with the other Macro-Ge languages.

9.5.5 Comparison with pair-wise n-gram comparison method

Finally, the pair-wise n-gram comparison method was applied to the same data, with chunking set to 1, and n-gram window 4. The resulting tree is shown in Figure 14.

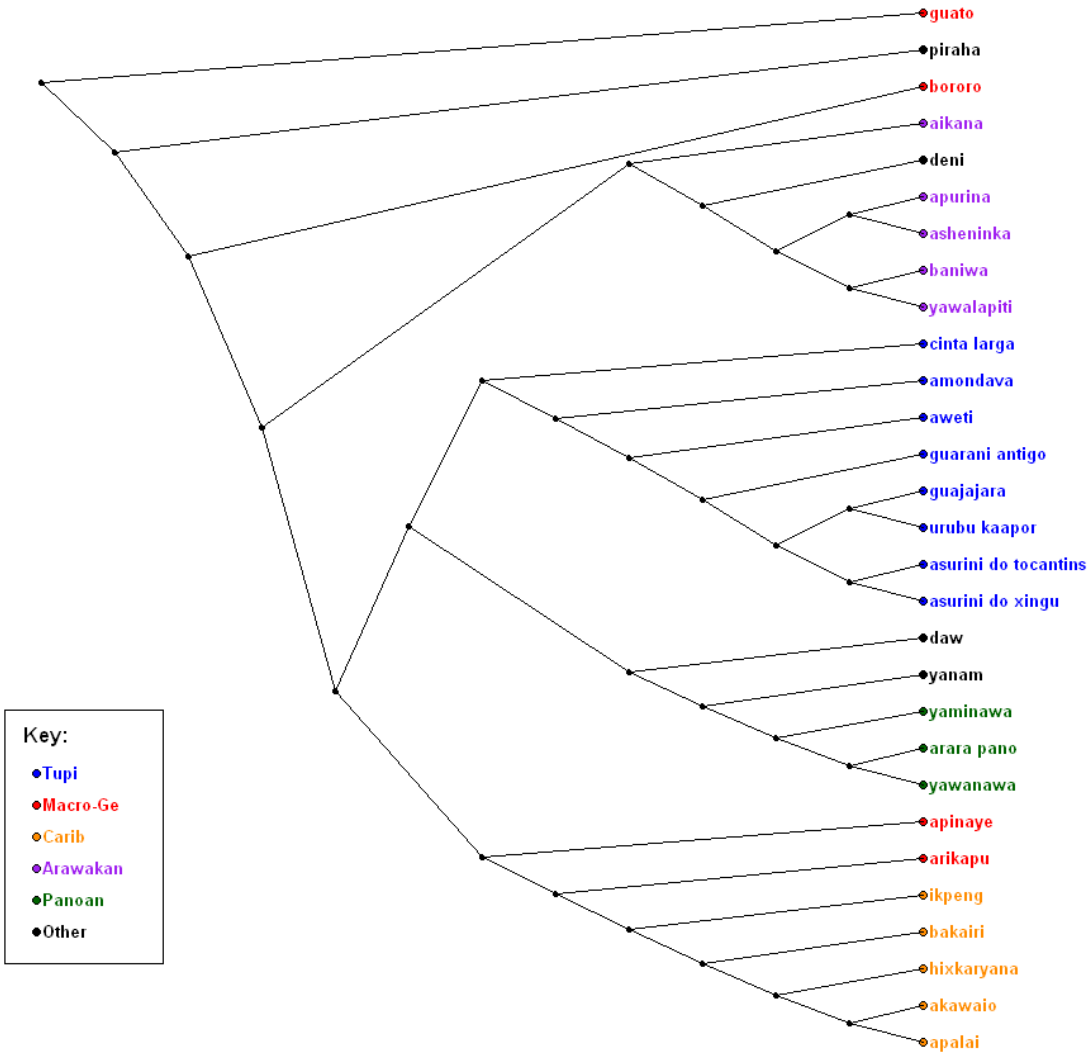

Figure 14: Tree produced by applying the pair-wise n-gram comparison method to the Brazilian data.

As with the other trees, the Macro-Ge family is the least well identified, but otherwise the tree fits extremely well with the generally accepted view.

10 Conclusions

Campbell (1997:207) says that "it is often by sheer chance that attention is turned to certain languages and not to others as being possible relatives of one another." This is because no method has existed to date which could be applied to large numbers of languages without detailed analysis having been carried out on those languages. The methods described here, based on n-gram comparison, provide a way to carry out analysis of languages whose relatedness is unknown.

As with any method, the results cannot be guaranteed to be correct. However, they can certainly be used as a way to cut through the volumes of data to find the few pairs of languages or language families that might deserve further attention. The majority of lexicostatistical methods simply cannot be applied in this way, as they are dependent on cognacy judgments, and in some cases (e.g. Heggarty, in McMahon and McMahon 2005:214-224), reconstructed forms.

The use of a detailed simulation of sound change is a major innovation of this thesis, and one which provides an unprecedented ability to test lexicostatistical methods scientifically.

On the basis of the tests described here, the following conclusions can be drawn:

- The pair-wise n-gram comparison method provides more accurate results than the standard n-gram comparison method or the initial phoneme comparison method, particularly with languages that have been subject to high levels of lexical and phonological borrowing. This was demonstrated using simulated language data.
- Using the full IPA transcription performs better, on average, than working with phonemes grouped by feature.
- Using 4-grams and 5-grams performs better, on average, than working with shorter n-gram windows.
- The n-gram comparison method depends on large volumes of text, although it also works well with the relatively short Swadesh lists.

Finally, although n-gram comparison is more complex to implement than the initial phoneme comparison method, it also allows much more detailed analyses of the data and has much greater scope for further development. I propose that further research could be carried out into the most effective ways to use the n-gram comparison and the pair-wise n-gram comparison method, and that these methods should be used to analyse the less well known languages of the world.

Appendix A - The Swadesh lists

The following are the meanings in the Swadesh 100 and 200 item lists. Swadesh selected 93 items from his original 200 item list, and added 7 new ones (shown in **bold**) to create his shorter 100 item list. Meanings excluded from the 100 item list are underlined.

all	dust	head	name	sharp	thou
and	ear	hear	narrow	short	three
animal	earth	heart	near	sing	throw
ashes	eat	heavy	neck	sit	tie
at	egg	here	new	skin	tongue
back	eye	hit	night	sky	tooth
bad	fall	hold (take)	nose	sleep	tree
bark	far	**horn**	not	small	turn
because	fat (grease)	how	old	smell	two
belly	father	hunt	one	smoke	vomit
big	fear	husband	other	smooth	walk
bird	feather	I	person (human being)	snake	warm (hot)
bite	few	ice	play	snow	wash
black	fight	if	pull	some	water
blood	fire	in	push	spit	we
blow	fish	kill	rain	split	wet
bone	five	**knee**	red	squeeze	what
breast	float	know	right (side)	stab (pierce)	when
breathe	flow	lake	right (true)	stand	where
burn	flower	laugh	river	star	white
child	fly	leaf	road (path)	stick	who
claw	fog	left (side)	root	stone	wide
cloud	foot	leg	rope	straight	wife
cold	four	lie	rotten	suck	wind
come	freeze	live	**round**	sun	wing
count	fruit	liver	rub	swell	wipe
cut	**full**	long	salt	swim	with
day	give	louse	sand	tail	woman
die	good	man	say	that	woods
dig	grass	many	scratch	there	worm
dirty	green	meat (flesh)	sea	they	ye
dog	guts	**moon**	see	thick	year
drink	hair	mother	seed	thin	yellow
dry	hand	mountain	sew	think	
dull (blunt)	he	mouth		this	

The (207 item) Swadesh lists for Slavic languages were obtained from Wiktionary.com: http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Slavic_languages

The (100 item) Swadesh lists for Brazilian languages were obtained from the following location: http://paginas.terra.com.br/educacao/GICLI/ListasEnglish.htm

An example of the 100 word lists for five of the languages is shown below:

English	**Aikanã**	**Akawaio**	**Amondava**	**Apalaí**	**Apinayé**
I	hisa	uːɾə	ɲihe	ɨwɨ	pa
you	hĩða	amərə	nehe	omoɾo	ka
we	satɛ	nja	nãne	ɨna	paʔtõʃ
this	hiba	seːɾə	koɾo	senɨ	já
that	kari	mərə	peɾo	mo	
what?	bari	ər	mãŋã	ota	mo
who?	tarai	ənɨk	ŋaɾa	onokɨ	mẽʔõ
not	hĩna	bɾa		pɨɾa	kət nẽ
all	amai	tamboɾo		emero	pijtã
many	taðaka	tuʔke	eʔhui	tuʔke	joʔto
one	amɛmɛ	tiginnə	oɲipeʔi	toiɾo	pɨtʃi
two	atuka	azaʔɾə	mõkõi	asakoro	atkru
big	tjabɨj	ege	hehãi	inũme	ratʃ
long	ũpɛ	kuzaŋ	ɨpɨ piɾuhu	mosa	rɨ
small	isiɛ̃	aigo	tʃuĩ	pisaɾaɾa	ŋri
woman	ðɛtja	atɨhpə	kũɲa	noʔpo	ni
man	kuɾɛða	waɾawok	kʷãmaʔe	oɾutua	bɨ
child		mɨɾe	taɨɾiʔga	poeto	kra
fish	ãti	moɾok	piɾa	kana	tɛp
bird	pijamamĩj	toroŋ	βiɾa	toɾono	kuwɐɲ
dog	ãrjya	kaiguzi	ɲaŋʷaɾa	kaikuʃi	rɔp
louse	kɨj	adan	kɨβa	azamo	ŋo
tree	wɛ	jəi	ɨβa	wewe	pĩ
seed	ðãw	te	haʔɨ̃ɲã	puʔtu	ʔɨ
leaf	widjdiʒi	eda	kaʔa	zaɾɨ	ʔo

English	**Aikanã**	**Akawaio**	**Amondava**	**Apalaí**	**Apinayé**
root	ðãpi	kara	ɨβapoa	mi	ʔare
bark of tree	ɛɗuɗu	pi:po		piʔpo	kʌ
skin	ɛɗuɗu	pi:po	pira	piʔpo	kʌ
flesh / meat	jẽ	puŋ	haʔo	punu	ĩ
blood	ĩ	mɨŋ	βɨʔɨ	munu	kamrɔ
bone	zu	eʔpɨ	kãŋã	zeʔpɨ	ʔi
egg	ðumɨ̃j	pəʔməj	hupiʔa	iʔmo	ŋrɛ
fat / grease	ðãjri	giajik	ikaβa		twəm
horn	kɨjɗɛ	iteʔrə	atia	ɾetɨ	pʌr
tail	wĩjdi	areʔna	βahaŋa	aroɨ	ʔamɨ
feather	ji	abiɾi	ipepoa	apori	ʔara
hair	ji	ʔpa	ʔabə	ʔpoti	kĩ
head	tinũpa	popo	akãŋã	puʔpy	krã
ear	kanĩðũ	wotahi	nãmia	pana	ʔamak
eye	kamuka	enu	akʷara	enu	nɔ
nose	kanãwã	enna	apɨ̃ŋã	euna	ʔĩakrɛ
mouth	kawa	mɨda	ɲurua	mɨta	akwa
tooth	mũj	ə	ahãŋã	ze	wa
tongue	waru	ja:ne	kõa	nuru	õʔtɔ
claw	iriðɨj	jenaʔpipə	po pea	emaʃipuʔtu	kɔp
foot	karɛtsa	pɨda	pɨa	pupu	par
knee	karɛmũ	zemuju	enepɨʔa	esekumu	kõn
hand	i:nɛ	emija	poa	ema	ʔĩkra
belly	katapa	wembo	eβega	waku	tu
neck	ɲɛnuũ	ɨʔmɨ	ɲura	pɨmɨ	
breast	tʃøtʃyʔĩ	manati	kãmã	manatɨ	
heart	tikʰikʰyʔi	ewaŋ	ɲaβeβuɲa	eano	mut
liver	i:ri	eɾi	pɨʔa	erɾ	ma
to drink	hu	eŋ	ɨʔu	enɨ	kõ
to eat	kaw	endaʔna	ʔu	otuku	krẽn
to bite	kaw	eka	huʔu	eseka	kaŋa
to burn	tarika	anuka	ahi	jaʔ	tʃey

English	Aikanã	Akawaio	Amondava	Apalaí	Apinayé
to see	apa	eneʔ	ɐpiek	ene	pumu
to hear	anapa	eda	ẽnu	eta	kuma
to know	arjo	iʔtu	kʷaha	waɾo	
to sleep	awã	eʔnumɨ	tʃiɾa	nɨkɨ	õt
to die	hĩmɛ̃	maʔta	mõnõ	oɾikɨ	tɨ
to kill	ta	wənə	ɲuka	etapa	kupĩ
to swim	sũ	etawa	ɲahɔg		re
to fly	tʃaw	waɾiwin	βeβe	pɐkã	tɔ
to walk		pinimɨ	ata	osenuʔ	mõ
to come	waɾɛ	jebɨ	ɾuɾi	oepɨ	te
to lie down	ty	epeɾenma	ʔãn	ataʃi	nõ
to sit down	dyry	eɾeuda	pɨ̃n	poɾoʔ	ɲɨ̃
to stand up	ɛwarjy	eʔmɨzaʔka	puʔam	owo	tʃa
to give	hiba	ɾeba	mõno	ekaɾo	ŋõ
to say	kjã	ka	ʔe	kaɾɨ	jarẽ
sun	jaðeɛrinɛʔi	wəi	kʷaɾa	ʃiʃi	mɨt
moon	ja	nuno	ɲahɨa	nuno	
star	jytɛ	siɾigu	ɲahɨtataʔia	ʃiɾikuato	kaŋɛti
water	hanɛ	tuna	ɨhɨa	tuna	ŋo
rain	hanɛ	konopo	ãmãnã	konopo	na
stone	haʒi	tək	ita	topu	kẽn
sand	hĩnũʔnuũ	sakow	ɨβɨtĩŋã	isawanɨ	
ground / earth	dy	noŋ	ɨβɨa	nono	pɨka
cloud	wirjyaʔi	katuɾu	ɨβaka	akuɾunu	kakrã
smoke	tʃyni	eʔsmokma	tatatĩŋã	eʃima	kũm
fire	hinɛ	watu	tata	apoto	kuwɨ
ash	ðũpapa	urumərəʔpə	tãnĩmuka	oɾuno	prə
path / road	ha	azauda	pehea	osema	pɾɨ
mountain	ui	wɨk	ɨbɨteɾa	ɨpɨ	
red	hadĩ	apiɾi	ɨβãŋahĩm	taʔpiɾe	kamrɛk
green	hørørø	soko	ɲakɨɾa	exuezume	
yellow	parari	sukupiju	tʃinahi	seweme	

English	**Aikanã**	**Akawaio**	**Amondava**	**Apalaí**	**Apinayé**
white	arara	aimorone	tĩŋãhĩm	karimutumano	ʔaka
black	vi	arikunan	ɲɨpiβahĩm	ʃinukutume	tɨk
night	ðũnɛ	goʔmamɨ	ɨpɨtũnã	koko	kamʌt
warm / hot	hãne	aʔnek	akoβ	aʃituneti	kaŋrɔ
cold	kjawɨj	komi	irotʃãŋahĩm	maʃi	akrɨ
full	jɛrɛwa	anesak	ha βaɲaβahĩm	peʔme	ʔipu
good	høʔã	wagɨ	ikatua	kure	mɛtʃ
new	ðamɛ	menaʔ	pɨahua	eʃisene	nɨw
round	urɛrɛpɛʔi	waitopan		pariʔme	
dry	hɛnɛkanɛʔi	aʔmunaga	ɨβirahĩm	tonore	ŋrʌ
name	kjawɨj	ezagɨ	era	eseti	itʃi

Appendix B - Coding

Experiments were carried out using scripts I wrote in the Ruby language (Thomas 2005). I chose this language because it is quick to develop in and has a large number of high-level string and vector handing functions built in. The graphical trees were produced using code written in Ruby with the Tk extension. In total, I wrote nearly 4,000 lines of Ruby code for this thesis.

One clear advantage of Ruby was its built-in Hash data structure, which I used to represent the n-gram vectors. This enormously simplified the development of the system, as it meant that very large but extremely sparse vectors could be stored in a reliable and efficient form, and without the need for development of new code or algorithms. In contrast, for example, Huffman (1998), coding in C++ was obliged to develop his own hashing mechanism, and furthermore to develop a somewhat complex chaining process to avoid issues caused by hash-table collisions (see Huffman 1998:151-155).

The following is a sample of the Ruby code used for this thesis. This script runs multiple t-tests and ANOVAs using a Monte Carlo methodology:

```
# statistical tests

require 'pearson'

def t_test_2_independent_groups(arr1,arr2)
  n = arr1.length
  ave1,ave2 = ave(arr1), ave(arr2)
  ssq1, ssq2, sx1, sx2 = 0.0, 0.0, 0.0, 0.0

  for i in 0..n-1 do
    ssq1 += (arr1[i] * arr1[i])
    ssq2 += (arr2[i] * arr2[i])
    sx1 += arr1[i]
    sx2 += arr2[i]
  end

  ss1 = ssq1 - ((sx1 * sx1) / n)
  ss2 = ssq2 - ((sx2 * sx2) / n)

  t = ((ave1 - ave2) / (Math.sqrt((ss1 + ss2) / (n * (n -1)))))
  return t
end

def t_test_paired_samples(arr1,arr2)
  n = arr1.length
  ssq = 0.0
  sx = 0.0
```

```
    for i in 0..n-1 do
      xd = arr2[i] - arr1[i]
      ssq += (xd * xd)
      sx += xd
    end

    ssd = ssq - ((sx * sx) / n)
    xd_bar = sx / n
    t = (xd_bar / (Math.sqrt(ssd / (n * (n-1)))))
  end

  def anova_1_way(arr1)
    # arr1 contains the data to be analysed.
    # arr1[0] should be a list of numbers for "subject" 0
    # arr1[i] is a list of numbers for "subject" i

    k = arr1.length
    total = 0.0
    for i in 0..k-1 do
      n = arr1[i].length
      sum = 0.0
      for j in 0..n-1 do
        sum += arr1[i][j].to_f
      end
      ssq = sum * sum
      total += (ssq / n)
    end

    all = arr1.flatten
    bigN = all.length
    sum = 0.0
    squares = 0
    for i in 0..bigN - 1 do
      sum += all[i].to_f
      squares += (all[i].to_f * all[i].to_f)
    end
    ssq = sum * sum

    ssb = total - (ssq / bigN)
    ssw = squares - total
  end

  def u_test(arr1,arr2)

  end

  def getFCriticalValues
    file = File.open("fcrit.values")

    arr = Array.new
    count = 0
    while line=file.gets do
      arr[count] = line.split.collect! {|x| x.to_f}
      count += 1
    end
    arr.transpose
  end

  def getTCriticalValues
    file = File.open("tcrit.values")
    arr = Array.new
    count = 1
    while line = file.gets do
      arr[count] = line.to_f
      count += 1
    end
```

```
  arr
end

def monteCarloTest(arr, iterations, test)
  # applies a statistical test to the data repeatedly,
  # taking a random sample from the largest set each time
  n1 = arr[0].length
  n2 = arr[1].length
  k = arr.length

  data = Array.new
  for i in 1..k-1 do
    data[i] = arr[i]
  end
  data[0] = Array.new

  # calculate the degrees of freedom:
  if test == "anova" then
    dfn = (arr.length - 1)
    dfd = (data.flatten.length - arr.length)
  end

  if test == "anova" then
    # calculate the critical alpha value:
    fcrit = getFCriticalValues
    alpha = fcrit[dfn][dfd]
  elsif test == "tp" then
    tcrit = getTCriticalValues
    alpha = tcrit[n2-1]
  elsif test == "ti" then
    tcrit = getTCriticalValues
    alpha = tcrit[n2 + n2 -1]
  end

  # if the lists are the same length, use all available data:
  if n1 == n2 then
    for i in 0..n1-1 do
      data[0][i] = arr[0][i]
    end
  end

  sig = 0
  iterations.times do |i|
    if n1 != n2 then
       # pick out random elements from arr[0] to populate data[0]
      for j in 0..n2-1 do
        data[0][j] = arr[0][rand(n1)]
      end
    end
    if test == "anova" then
      f = anova_1_way(data)
      if f > alpha then sig += 1 end
      if iterations == 1 then print "f = " + f.to_s + "\n" end
    elsif test == "tp" then
      t = t_test_paired_samples(data[1], data[0])
      if t > alpha then sig += 1 end
      if iterations == 1 then print "t = " + t.to_s + "\n" end
    elsif test == "ti" then
      t = t_test_2_independent_groups(data[1], data[0])
      if t > alpha then sig += 1 end
      if iterations == 1 then print "t = " + t.to_s + "\n" end
    end
  end

  print sig.to_s + " out of " + iterations.to_s + " trials were significant\n"
  print "critical alpha value = " + alpha.to_s + "\n"
```

```
end

def readNumbers(filename)
  file = File.open(filename,"r")
  arr = Array.new
  while line=file.gets
    arr.push(line.to_f)
  end
  arr
end

if ARGV.length < 3 then
  print "usage: " + $0 + " [options] iterations filename1 filename2 [...
filenamen]\n"
  print "options:\n"
  print "-tp    : paired samples t-test\n"
  print "-ti     : independent samples t-test\n"
  print "-a     : 1-way anova (default)\n"
  Process.exit()
end


anova = true
ttest_ind = false
ttest_paired = false
while ARGV[0][0].chr == "-"
# a while loop so users can combine options:
  if ARGV[0] == "-tp" then
    anova = false
    ttest_paired = true
  end

  if ARGV[0] == "-ti" then
    anova = false
    ttest_ind = true
  end

  ARGV.slice!(0)
end

numIterations = ARGV[0].to_i
ARGV.slice!(0)

arr = Array.new
for i in 0..ARGV.length-1 do
  arr[i] = readNumbers(ARGV[i])
end

if anova then
  monteCarloTest(arr,numIterations,"anova")
end

if ttest_paired then
  monteCarloTest(arr,numIterations,"tp")
end

if ttest_ind then
  monteCarloTest(arr,numIterations,"ti")
end
```

Appendix C - References

Barbançon, F., Warnow, T., Evans, S., Ringe, D., Nakhleh, L. (2007) *An experimental study comparing linguistic phylogenetic reconstruction methods.* U.C. Berkeley Department of Statistics Technical Report No. 732.

Baumhoff, M., Olmsted, D. (1963) Palaihnihan: radiocarbon support for Glottochronology. *American Anthropologist*, New Series, Vol. 65, No. 2, Apr 1963, 278-284.

Bergsland, K., Vogt, H. (1962) On the Validity of Glottochronology. *Current Anthropology*, Vol. 3, No. 2, Apr., 1962, 115-153.

Black, P. (2007) Lexicostatistics with massive borrowing: The case of Jingulu and Mudburra. *Australian Journal of Linguistics*, Vol. 27, No. 1. 63-71.

Borland, C. (1982) How Basic is "Basic" vocabulary? *Current Anthropology*, Vol. 23, No. 3. 315-316.

Bryant, D. (2006) Radiation and network breaking in Polynesian Linguistics. In P. Forster & C. Renfrew (eds), *Phylogenetic Methods*. Cambridge: McDonald Institute for Archaeological Research. 111-118.

Bynon, T. (1977) *Historical Linguistics*. Cambridge: Cambridge University Press.

Campbell, L. (1997) *American Indian Languages: The historical linguistics of native America*. Oxford: Oxford University Press.

Campbell, L. (1998) *Historical Linguistics*. Edinburgh: Edinburgh University Press.

Crowley, T. (1992) *An Introduction to Historical Linguistics*. Oxford: Oxford University Press.

Coppin, B. (2008) A critical review of the techniques used in the lexicostatistical analysis of language relatedness. Unpublished manuscript. University of Cambridge.

Damashek, M. (1995) Gauging similarity with n-grams: language-independent categorization of text. *Science*, New Series, Vol. 267. No. 5199. 843-848.

Dunn, M., Terrill, A., Reesink, G., Foley, R., Levinson, S. (2005) Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science*, Vol. 309. 2072-2075.

Dyen, I., Kruskal, J., Black, P. (1992) An Indo-European classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* vol 82, part 5, 1992.

Ellison, M. & Kirby, S. (2006) Measuring Language Divergence by Intra-Lexical Comparison. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 273–280, Sydney, 2006.

Embleton, S. (1986) *Statistics in Historical Linguistics*. Bochum: Studienverlag Brockmeyer.

Felsenstein, F. (2004) *Inferring Phylogenies*. Sunderland, Massachusetts: Sinauer Associates, Inc.

Forster and Toth (2003) Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Science* Vol. 100 No. 15. 9079-9084

Gordon, Raymond G. (ed.), (2005). *Ethnologue: Languages of the World*. Dallas, Texas: SIL International.

Gray, R., Atkinson, Q. (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, Vol. 426, 27 November 2003. 435-439.

Gudschinsky, S. (1956) The ABCs of Lexicostatistics (Glottochronology). *Word*, Vol. 12, No. 2, 175-210.

Guy, J. (1980) *Experimental Glottochronology: Basic Methods and Results.* Pacific Linguistics Series B, No. 75.

Hale, M. (2007) *Historical Linguistics: Theory and Methods*. Oxford: Blackwell Publishing.

Hartman, L. (2003) Modeling Phonological Change. Web page, accessed 5th June 2008. http://mypage.siu.edu/lhartman/phono/modeling.htm.

Heggarty, P. (2000) Quantifying change over time in phonetics. In C. Renfrew, A. McMahon, L. Trask (eds), *Time depth in historical linguistics*. Cambridge: The McDonald Institute for Archaeological Research. 531-562.

Hirsch, D. (1954) Glottochronology and Eskimo and Eskimo-Aleut Prehistory. *American Anthropologist*, New Series, Vol. 56, No. 5, Part 1 (Oct., 1954), 825-838.

Hock, H., Joseph, B. *Language history, language change and language relationship.* Berlin: Mouton de Gruyter.

Huffman, S. *The Genetic Classification of languages by n-gram analysis: a computational technique*. Unpublished PhD dissertation, Georgetown University.

Hymes, D. (1971) Lexicostatistics and Glottochronology in the nineteenth century (with notes towards a general history). Offprint from I. Dyen (ed), *Lexicostatistics in Genetic Linguistics, Proceedings of the Yale Conference.* Yale University, April 3-4 1971. Published 1973. The Hague:Mouton.

Kessler, B. (2001) *The Significance of Word Lists.* Stanford: CSLI Publications.

Kessler, B., Lehtonen, A. (2007) Word Similarity Metrics and Multilateral Comparison. *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology* (6–14). Stroudsburg PA: Association for Computational Linguistics.

Longobardi, G. (2003) Methods in Parametric Linguistics and Cognitive History. Unpublished manuscript. University of Trieste.

Maddieson, I. (1984) *Patterns of sounds*. Cambridge: Cambridge University Press.

McMahon, A., McMahon, R.(2005) *Language Classification by Numbers*. Oxford: Oxford University Press.

Meillet, A. (translated by Gordon B. Ford Jr.) (1970; originally 1928) *The Comparative Method in Historical Linguistics*. Paris: Librairie Honore Champion.

Nicholls, G. and Gray, R. (2006) Quantifying uncertainty in a stochastic model of vocabulary evolution. In P. Forster & C. Renfrew (eds), *Phylogenetic Methods.* Cambridge: McDonald Institute for Archaeological Research, 161-171.

Nichols, J. (2006) Quasi-cognates and lexical type shifts: rigorous distance measures for long-range comparison. In P. Forster & C. Renfrew (eds), *Phylogenetic Methods*. Cambridge: McDonald Institute for Archaeological Research, 57-65.

Pagel, M., Atkinson, Q. & Meade, A. (2007) Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*: 449, 717-720.

Pagel, M. and Meade, A. (2006) Estimating rates of lexical replacement on phylogenetic trees of languages. In P. Forster & C. Renfrew (eds), *Phylogenetic Methods.* Cambridge: McDonald Institute for Archaeological Research, 173-181.

Paulston, C. and Peckham, D. (1998) *Linguistic Minorities in Central and Eastern Europe.* Clevedon: Multilingual Matters Limited.

Poser, W. (2004) *Gray and Atkinson - Use of Binary Characters.* http://itre.cis.upenn.edu/%7Emyl/languagelog/archives/000832.html (accessed 8th January 2007).

Rigon, G. (2007) Towards the automatic detection of syntactic borrowing on a parametric dataset. Unpublished manuscript. University of Trieste.

Ringe, D. (1992) On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society*, Volume 82, Part 1, 1992. Philadelphia.

Ringe, D., Warnow, T., Taylor, A. (2002) Indo-European and Compuational Cladistics. *Transactions of the Philological Society*, Vol. 100 No. 1. 59-129.

Ross, A. (1950) Philological probability problems. *Journal of the Royal Statistical Society.* Series B, 12: 19-59.

Sankoff, D., Sankoff, G.(1976) Wave versus Stammbaum explanations of lexical similarities. In I. Dyen, G. Jucquois (eds), *Lexicostatistics in Genetic Linguistics II*. Proceedings of the Montreal Conference, May 19-20, 1973. 29-41.

Scannell, K. (2004). Web page, accessed 5th June 2008. http://borel.slu.edu/crubadan/index.html

Shannon, C. (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal*, Vol. 27, 379–423 & 623–656.

Shibatani, M. (1973) The Role of Surface Phonetic Constraints in Generative Phonology. *Language*, Vol. 49, No. 1. 87-106.

Sommerstein, A. (1974) On Phonotactically motivated rules. *Journal of Linguistics* Vol. 10. 71-94.

Stuart, G., Moffett, K and Baker, S. (2002) Integrated Gene and Species Phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, Vol. 18, No.1. 100-108.

Sussex, R., Cubberley, P. (2006) *The Slavic Languages*. Cambridge: Cambridge University Press.

Swadesh, M. (1950) Salish Internal Relationships. *International Journal of American Linguistics*, Vol. 16, No. 4. (Oct., 1950), 157-167.

Swadesh, M. (1951) Diffusional Cumulation and Archaic Residue as Historical Explanations. *Southwestern Journal of Anthropology*, Vol. 7, No. 1. 1-21.

Swadesh, M. (1953) Comment on Hockett's Critique. *International Journal of American Linguistics*, Vol. 19, No. 2. 152-153.

Swadesh, M. (1954) Symposium: Time Depths of American Linguistic Groupings. *American Anthropologist*, New Series, Vol. 56, No. 3. (Jun., 1954), 361-377.

Swadesh, M. (1955) Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, Vol. 21, No. 2. (Apr., 1955), 121-137.

Teeter, K. (1963) Lexicostatistics and Genetic Relationship. *Language*. Vol. 39, No. 4. (Oct. - Dec., 1963), 638-648.

Teeter, K. (1965) Remarks on Diebold, 'A Control Case for Glottochronology'. *American Anthropologist*, New Series, Vol. 67, No. 6, Part 1 (Dec 1965), 1522-1524.

Thomas, D. (2005) *Programming Ruby: The pragmatic programmers' guide.* Raleigh, North Carolina: The Pragmatic Programmers, LLC.

van den Bosch, A., Content, A., Daelemans, W. and de Gelder, B. (1994) Measuring the complexity of writing systems. *Journal of Quantitative Linguistics*, Vol. 1. No. 3. 178-188.