Identification of Cognates and Recurrent Sound Correspondences in Word Lists

Grzegorz Kondrak

Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada. E-mail: kondrak@cs.ualberta.ca.

ABSTRACT. *Identification of cognates and recurrent sound correspondences is a component of two principal tasks of historical linguistics: demonstrating the relatedness of languages, and reconstructing the histories of language families. We propose methods for detecting and quantifying three characteristics of cognates: recurrent sound correspondences, phonetic similarity, and semantic affinity. The ultimate goal is to identify cognates and correspondences directly from lists of words representing pairs of languages that are known to be related. The proposed solutions are language independent, and are evaluated against authentic linguistic data. The results of evaluation experiments involving the Indo-European, Algonquian, and Totonac language families indicate that our methods are more accurate than comparable programs, and achieve high precision and recall on various test sets. The results also suggest that combining various types of evidence substantially increases cognate identification accuracy.*

RÉSUMÉ. *L'identification de mots apparentés et des correspondances de sons récurrents intervient dans deux des principales tâches de la linguistique historique: démontrer des filiations linguistiques et reconstruire l'histoire des familles de langues. Nous proposons des méthodes de détection et de quantification de trois caractéristiques des mots apparentés: les correspondances de sons récurrents, la ressemblance phonétique et l'affinité sémantique. Le but ultime est d'identifier les mots apparentés et les correspondances directement à partir de listes de mots représentant des paires des langues dont la filiation est connue. Les solutions proposées sont indépendantes des langues traitées et sont évaluées sur des données linguistiques réelles. Les résultats d'expériences impliquant des langues indo-européennes, algonquines et des langues de la famille du totonaque indiquent que nos méthodes sont plus précises que des programmes comparables et d'atteignent une haute précision et un haut taux de rappel sur des ensembles de test. Les résultats suggèrent également que la combinaison de divers types d'indices augmente grandement la justesse de l'identification des mots apparentés.*

KEYWORDS: *cognates, sound correspondences, word lists, diachronic studies, proto-languages*

MOTS-CLÉS : *cognats, correspondances phonétiques, études diachroniques, proto-langages*

1. Introduction

Identification of cognates and recurrent sound correspondences is a component of two principal tasks of historical linguistics: demonstrating the relatedness of languages and reconstructing the histories of language families. Genetically related languages originate from a common proto-language. In the absence of historical records, proto-languages have to be reconstructed from cognates — reflexes of proto-words that survive in some form in the daughter languages. Sets of cognates regularly exhibit recurrent sound correspondences. Thus, cognates and recurrent sound correspondences provide strong evidence of a common origin of languages.

Over the last two hundred years, historical linguists have developed *the comparative method* of language reconstruction, which involves identification of cognates and recurrent sound correspondences. Numerous languages from around the world have been shown to be related by applying the comparative method. In particular, scholars have reconstructed well over a thousand roots of Proto-Indo-European, the hypothetical ancestor of most European and Indian languages. However, language reconstruction is an extremely time-consuming process that has yet to be accomplished for many language families. For example, Greenberg (1993) calls the task of reconstructing proto-languages of his proposed eleven Amerind subgroupings "superhuman," and estimates that it could take him "several centuries of effort." Nevertheless, most linguists insist on corroborating claims of relatedness with a list of recurrent sound correspondences.

Since the task of the identification of cognates and correspondences involves detecting regularities in large amounts of data, it is natural to ask whether it can be performed by a computer program. In this article, we propose methods for detecting and quantifying three characteristics of cognates: recurrent sound correspondences, phonetic similarity, and semantic affinity. The ultimate goal is to identify cognates and correspondences directly from lists of words representing pairs of languages that are known to be related. Our general approach is to combine novel algorithms developed specifically for the task at hand with algorithms adapted from bioinformatics and natural language processing. The proposed solutions are language independent and are evaluated against authentic linguistic data.

The computer programs that implement the methods described in the following sections are not meant to replace historical linguists. On the contrary, they are intended as aids for exploratory analysis of little-studied languages, and their output must be critically examined by well-informed experts. Accordingly, they are evaluated against other programs developed for the same purpose, rather than against human performance. Furthermore, the methods are designed to be applied to pairs of languages whose genetic relationship is beyond doubt. They are unsuitable for determining whether two languages are related, although they could potentially be used to furnish additional evidence for supporting claims of relatedness. The programs are implemented in C++ and are freely available to interested researchers. They have been already applied in projects involving several diverse language families.

Although our primary focus is historical linguistics, the methods that we propose have a wider scope. Parallel bilingual corpora (*bitexts*) have been increasingly important in statistical natural language processing. Cognates have been employed for a number of bitext-related tasks, including sentence alignment (Melamed, 1999), word alignment (Tiedemann, 1999), and inducing translation lexicons (Mann and Yarowsky, 2001). The line of research that this article represents has already resulted in applications in such diverse areas as statistical machine translation (Kondrak *et al.*, 2003) and the identification of confusable drug names (Kondrak and Dorr, 2006). In the long run, such applications may prove even more important than the original linguistic motivation of the research that led to them. However, the language reconstruction framework is particularly well-suited for formulating the driving problems and for testing the proposed solutions.

The rest of the article has the following structure: Section 2 defines the task in more detail. Section 3 discusses related work. Sections 4, 5, 6, and 7 are devoted to four sources of evidence for cognation: phonetic similarity, simple (one-to-one) correspondences, complex (many-to-many) correspondences, and semantic similarity. Section 8 describes a method of combining those types of evidence into a single score. Sections 9, 10, and 11 describe the results of evaluation experiments involving the Indo-European, Algonquian, and Totonac-Tepehua language families, respectively. Section 12 concludes the article.

2. Background

In the narrow sense used in historical linguistics, cognates are words in related languages that have gradually developed from the same ancestor word. An example of a cognate pair is French *lait* and Spanish *leche*, both of which come from Proto-Romance *lacte*. In other contexts, the term is often used more loosely, denoting words in different languages that are similar in form and meaning, with no distinction between borrowed and genetically related words; for example, English *sprint* and the Japanese borrowing *supurinto* are considered cognate, even though these two languages are unrelated. In this article, we adhere to the strict sense of the term “cognate,” which excludes borrowings.

Due to their common origin, cognates often sound alike and have similar meaning. However, with time, cognates often acquire very different phonetic shapes. For example, English *hundred*, French *cent*, and Polish *sto* are all descendants of Proto-Indo-European **k̂ṃtom* (an asterisk denotes a reconstructed form). The semantic change can be no less dramatic; for example, English *guest* and Latin *hostis* ‘enemy’ are cognates even though their meanings are diametrically different. On the other hand, phonetic similarity of semantically equivalent words can also be due to other factors, such as lexical borrowing (direct or from a third language), onomatopoeia, nursery words, and chance resemblance.

Meaning	English	Latin	Cognate	Correspondences
‘to fight’	f a j t	p u g n ā	-	
‘foam’	**f** ō **m**	s **p** u **m**	√	**f:p**, **m:m**
‘foot’	**f** u **t**	**p** e **d**	√	**f:p**, **t:d**
‘heart’	**h** a **r** **t**	**k** o **r** **d**	√	**h:k**, **r:r**, **t:d**
‘horn’	**h** ɔ **r** **n**	**k** o **r** **n**	√	**h:k**, **r:r**, **n:n**
‘hot’	h a t	k a l i d	-	
‘to rub’	r ə b	f r i k ā	-	
‘to scratch’	s k r æ ʧ	s k a b e	-	
‘three’	**θ** **r** ī	**t** **r** e	√	**θ:t**, **r:r**
‘tooth’	**t** ū **θ**	**d** e n **t**	√	**t:d**, **θ:t**

Table 1. *A sample of word stems from a bilingual word list with cognates and consonant correspondences identified*

In the past, languages were often grouped in families on the basis of similarity of basic vocabulary. Nowadays, most linguists insist on corroborating the claims of relatedness with a list of sound correspondences that recur in cognates. For example, sound correspondences between English and Latin include **f:p**, **t:d**, **h:k**, and **n:n** (Table 1). Sound correspondences in cognates are preserved over time in some form thanks to the regularity of sound changes, which normally apply to sounds in a given phonological context across all words in the language. Although apparent sound correspondences may sometimes arise in sets of unrelated words, correspondences are generally considered to provide much stronger evidence for cognation than phonetic similarity.

The tasks of the identification of cognates and the identification of recurrent sound correspondences are intertwined. In order to make reliable judgments of cognation, it is necessary to know what the correspondences are. However, the correspondences can only be extracted from word pairs that are genuine cognates.

Depending on the kind of data, the task of cognate identification can be defined on three levels of specificity:

1) Given a pair of words, such as English *snow* and German *Schnee*, compute a relative score reflecting the likelihood that they are cognate.

2) Given a list of word pairs matched by meanings, such as the one in Table 1, rank the pairs according to the likelihood that they are cognate.

3) Given a pair of vocabulary lists, such as the one in Table 2, produce a ranked list of candidate cognate pairs.

A phonetic similarity measure can be computed for any pair of words in isolation (levels 1, 2, and 3), but the determination of the recurrent sound correspondences requires a list of related words (levels 2 and 3), while a semantic measure is applicable

Cree	Gloss	Ojibwa	Gloss
āniskōhōčikan	string of beads	*āšikan*	dock, bridge
asikan	sock, stocking	*anakaʔēkkw*	bark
kamāmakos	butterfly	*kipaskosikan*	medicine
kostāčīwin	terror, fear	*kottāčīwin*	fear, alarm
misiyēw	large partridge, hen	*mēmīkwanʔ*	butterfly
namēhpin	wild ginger	*misissē*	turkey
napakihtak	board	*namēpin*	sucker
tēhtēw	green toad	*napakissakw*	plank
wayakēskw	bark	*tēntē*	very big toad

Table 2. *Samples from vocabulary lists representing two related languages*

when words are accompanied by glosses or some other representation of their meaning (level 3 only).

The comparative method is the technique used by linguists to reconstruct proto-forms of the parent language by examining cognates in its daughter languages (Trask, 1996). It consists of several stages. First, words with similar meanings are placed side by side. Those pairs that exhibit some phonological similarity are identified as putative cognates. Next, the cognates are aligned by pairing related phonetic segments, and analyzed in order to find systematic correspondences. A proto-phoneme or a proto-allophone is posited for each established correspondence. The proto-forms that gave rise to the identified cognate sets are then reconstructed. The resulting phonological system of the proto-language is adjusted in order to conform to general linguistic principles, and to take into account available data from ancestral and more distantly related languages. Naturally, the results of later steps can be used to refine the judgments made in earlier ones. Although the term “comparative method” suggests an algorithm that could be directly implemented on a computer, it is more a collection of heuristics, which involve intuitive criteria and broad domain knowledge.

3. Related work

Since most sound changes are regular, it is relatively straightforward to design a derivation program that takes advantage of this regularity in order to simulate evolution of languages. Such programs have been constructed for tracing forms from Latin to Spanish (Eastlack, 1977; Hartman, 1981), Latin to French (Burton-Hunter, 1976), and Proto-Indo-European to Russian (Smith, 1969). They derive later forms by the application of a set of phonological changes to earlier forms. Raman *et al.* (1997) use derivation programs to develop distance measures between parent and daughter languages in Chinese dialects.

Proceeding in the other direction, it is possible to derive proto-forms from the modern forms on the basis of recurrent sound correspondences provided by the user. By identifying identical proto-forms back-generated from various modern forms, comparative dictionaries have been constructed for the Algonquian (Hewson, 1974) and the Yuman family of languages (Johnson, 1985). The *Reconstruction Engine* (Lowe and Mazaudon, 1994) is a more general proposal designed to aid the historical linguist in reconstruction work. It consists of a suite of programs that not only generates reflexes from the provided proto-forms, but establishes cognate sets together with reconstructions by processing entire lexicons of related languages.

The common characteristic of the three approaches is their dependence on previously determined recurrent sound correspondences. Unfortunately, the determination of sound correspondences is one of the most challenging steps of the reconstruction process. Complete tables of correspondences can be constructed for well-studied language families on the basis of previously identified cognate sets, but are not available for many African and American languages, especially in the cases where the relationship between languages has not been adequately proven. A linguist whose job is to retrace the development of a language family may have only word lists of modern forms at her disposal. In all but a handful of cases, there are no historical records that demonstrate the form of the proto-language.

A few proposals have been aimed at meeting the challenge of the automatic discovery of cognates and correspondences from word lists. Kay (1964) presents an interesting attempt to formalize a large part of the comparative method in terms of propositional logic. The criterion that guides the search for correspondences is the minimization of the number of proto-phonemes necessary to account for the input data. However, the method is computationally impractical, and its ability to handle noisy data is doubtful. Damerau (1975) describes an algorithm for finding recurrent correspondences in word lists. Word pairs that are completely covered by the correspondences are classified as cognates. Guy (1994) outlines a correspondence-based algorithm for identifying cognates in bilingual word lists. He presents no quantitative evaluation of the method on authentic language data, but the program COGNATE that implements the algorithm is publicly available for testing. The Covington (1996) proposal is limited to finding the optimal alignment of cognate pairs using depth-first search and a phoneme-class distance function. Oakes (2000) describes a set of programs named JAKARTA that together perform several steps of the comparative method, from the determination of recurrent correspondences in word lists to the actual reconstruction of the proto-forms. The weak point of the proposal is that it was developed and evaluated on the same set of word lists representing four Indonesian languages. I evaluate the performance of COGNATE and JAKARTA in Sections 9 and 10.

The estimation of the likelihood of historical connection between languages is a task related to the one at hand. In order to compute the probability that the correlation between languages is statistically significant, Baxter and Ramer (2000) and Oswalt (1998) employ measures based on phonetic similarity, while Ringe (1998) and

Kessler (2001) concentrate on recurrent sound correspondences. The bilingual word lists employed in that line of research are of the same kind as the ones used for the experiments described in this article. However, those methods are aimed at languages whose relatedness has not been yet firmly established, and do not provide directly verifiable evidence in the form of identified cognates and correspondences.

Some approaches to cognate identification focus on goals unrelated to historical linguistics. Tiedemann (1999) investigates automatic construction of weighted string similarity measures from bitexts, and Mann and Yarowsky (2001) consider automatic induction of translation lexicons between related languages. Both methods implicitly determine and employ correspondences. In their paper on back-transliteration, Knight and Graehl (1998) compute symbol-mapping probabilities between English and Japanese. It is possible to view the sound pairs with the highest probabilities as the strongest correspondences between the two languages.

4. Orthographic and phonetic similarity

The approaches to measuring word similarity (or word distance) can be divided into two groups. The orthographic approaches disregard the fact that alphabetic characters express actual sounds, employing a binary identity function on the level of character comparison. The phonetic approaches, on the other hand, attempt to take advantage of the phonetic characteristics of individual sounds in order to estimate their similarity. The words are assumed to be represented in a phonetic or phonemic notation. Intuitively, complex phonetic algorithms should be more accurate than simple, "orthographic" measures. By applying various methods to the specific task of cognate identification, their relative performance can be objectively evaluated.

4.1. *The orthographic approaches*

One of the simplest cognate identification approaches was proposed by Simard *et al.* (1992). They consider two words to be cognate (in the broad sense of the word) if they are at least four characters long and their first four characters are identical. This approach can be generalized by defining a PREFIX coefficient which returns values in the $[0, 1]$ range. PREFIX is computed by dividing the length of the longest common prefix by the length of the longer word. For example, PREFIX(*colour*,*couleur*) = $\frac{2}{7} \simeq 0.29$ because their longest common prefix is *co*-.

Dice's similarity coefficient, originally developed for the comparison of biological specimens, was first used to compare words by Adamson and Boreham (1974). It is based on the notion of a bigram — an ordered pair of characters. Dice's coefficient is determined by the ratio of the number of shared character bigrams to the total number of bigrams in both words. For example, *colour* and *couleur* share three bigrams (*co*, *ou*, and *ur*), so their Dice's coefficient is $\frac{2 \times 3}{11} \simeq 0.55$.

Melamed (1999) detects orthographic cognates by thresholding the Longest Common Subsequence Ratio (LCSR). A common subsequence is a sequence composed of units appearing in both words, respecting their order, but not necessarily contiguous. The LCSR of two words is computed by dividing the length of their longest common subsequence by the length of the longer word: For example, LCSR(*colour*,*couleur*) = $\frac{5}{7} \simeq 0.71$, as their longest common subsequence is *c-o-l-u-r*. LCSR is closely related to *edit distance*, which is defined as the minimum number of substitutions, insertions, and deletions necessary to convert one word into another (Wagner and Fischer, 1974). If the cost of a substitution is set at twice the cost of an insertion/deletion, the length of the longest common subsequence between two words can be computed directly from their edit distance.

4.2. *The phonetic approaches*

The phonetic approaches are usually based on the decomposition of phonemes into vectors of phonetic features. Both Kessler (1995) and Nerbonne and Heeringa (1997) developed such methods for the task of measuring phonetic distance between dialects.

JAKARTA is a phonetic-based approach developed specifically for the purpose of cognate identification (Oakes, 2000). Two words are deemed to be cognate if their edit distance is below a certain threshold. The threshold was established by the analysis of the distances between cognate and non-cognate pairs in four Indonesian word lists. The phonetic characteristics of sound are stored by means of just three features: place, manner, and voicing, each of which has several possible values. Thus, distinct phonemes can have identical feature assignments. The similarity between phonetic segments is estimated by checking the identity of the feature values only; there is no notion of the relative distance between various places or manners of articulation.

4.3. *ALINE*

ALINE (Kondrak, 2000) was originally developed for aligning corresponding phonemes in cognate pairs, which is an essential step in the comparative method of language reconstruction. However, since it chooses the optimal alignment on the basis of a similarity score, it can also be used for computing phonetic similarity.

The principal component of ALINE is a function that calculates the similarity of two phonemes. Phonemes are expressed in terms of binary or multi-valued phonetic features. For example, the phoneme [n], which denotes a voiced alveolar nasal stop, has the following feature values: *Place = 0.85*, *Manner = 0.6*, *Voice = 1*, and *Nasal = 1*, with the remaining features set to 0. In order to compute the phonetic distance between two phonemes, the differences between their numerical values for each feature are multiplied by the feature's salience weight, and the resulting values are summed up. The similarity score is then calculated by subtracting the distance from the maximum score between two phonemes. For the purpose of emphasizing consonant correspondences, the similarity score is further decreased if one or both of the phonemes are vowels (vowel penalty).

The feature set contains the following features: *Place*, *Manner*, *Voice*, *Syllabic*, *Nasal*, *Retroflex*, *High*, *Lateral*, *Aspirated*, *Back*, *Round*, and *Long*. A special feature *Double*, which has the same possible values as *Place*, indicates the second place of articulation. The above feature set is sufficient to account for phonemic contrasts in many languages, and can be extended to cover other languages, if necessary.

The numerical feature values reflect the distances between vocal organs during speech production, and are based on the values reported by (Ladefoged, 1975, pages 258–259). For example, the feature *Manner*, which, roughly speaking, refers to the degree of airstream opening in the vocal tract during phoneme articulation, can take any of the following seven values: *stop* = 1.0, *affricate* = 0.9, *fricative* = 0.8, *approximant* = 0.6, *high vowel* = 0.4, *mid vowel* = 0.2, and *low vowel* = 0.0.

An important component of ALINE's feature system is the notion of the salience weights that represent the relative importance of each feature. The principal features, *Place* and *Manner*, are assigned much higher salience weights than less important features like *Aspirated* and *Round*. The default salience values were established by trial and error on a set of phoneme-aligned cognate pairs from various related languages.

The overall similarity score and optimal alignment of two words are computed by a dynamic programming algorithm (Wagner and Fischer, 1974). The total score is the sum of individual similarity scores between pairs of phonemes in the optimal alignment. A constant insertion/deletion penalty is applied for each unaligned phoneme. ALINE incorporates a number of extensions to the basic dynamic programming, which have been proposed primarily to address issues in DNA alignment, but are also applicable in the context of computing phonetic word similarity. The extensions include: retrieving a set of best alignments (Myers, 1995), local and semiglobal alignment (Smith and Waterman, 1981), and additional edit operations (Oommen, 1995).

The similarity score returned by ALINE is normalized by dividing it by the length of the longer word multiplied by the maximum possible similarity score between two phonemes, so that it falls in the range $[0, 1]$. Because it uses similarity rather than distance, the score assigned to two identical words is not a constant, but depends on the length and content of the words.

The feature system of ALINE is highly dynamic because the phonetic similarity values between phonemes can be modified by changing both feature salience weights and numerical values within features. Additional parameters include the maximum score between two phonemes, the insertion/deletion penalty, and the vowel penalty. The parameters have default settings for the cognate matching task, but can be manually optimized (tuned) on training data sets that include both cognates and unrelated word pairs. A complete description of ALINE can be found in (Kondrak, 2003b).

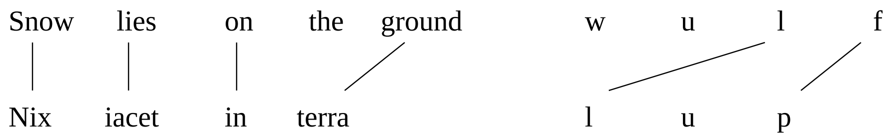

Figure 1. *The similarity of word alignment in bitexts and phoneme alignment between cognates*

5. Determination of simple sound correspondences

The approaches described in the previous section can be used for the identification of cognates on the basis of phonetic or orthographic similarity. However, such algorithms align one word pair at a time, and have no ability to generalize from larger data sets. Most linguists believe that recurrent sound correspondences provide a more reliable evidence of cognation. For example, the English verb *have* is not cognate with Latin *habere* ‘to have,’ as implied by the phonetic and semantic similarity, but rather with *capere* ‘to catch.’ This follows from the well-known Grimm’s Law, which specifies that English [h] regularly corresponds to Latin [k]. The presence of correspondences is what distinguishes cognates from loan words and chance resemblances.

A strong similarity between the task of matching phonetic segments in a pair of cognate words, and the task of matching words in two sentences that are mutual translations has been noticed before: “there is an interesting parallel to be drawn between the comparison of words which is made in the comparative method and the comparison of sentences made in the study of translation. The object of both enterprises is to specify transformations of one set of strings into another.” (Kay, 1964, page iii) The consistency with which a word in one language is translated into a (not necessarily cognate) word in another language is mirrored by the consistency of sound correspondences. The former is due to the semantic equivalence, while the latter follows from the principle of the regularity of sound change. Thus, it should be possible to use similar techniques for both tasks. Figure 1 illustrates the analogy between the recurring correspondences of translations in bilingual corpora and sounds in cognate lists.

Statistical machine translation is the method of generating translation systems automatically from large bitexts (Koehn, 2009). The idea is to combine a *language model*, which assigns a probability to every sentence in the target language, with a *translation model*. A translation model approximates the probability that two sentences are mutual translations by computing the product of the probabilities that each word in the target sentence is a translation of some source language word. A model of translation equivalence that determines the word translation probabilities can be *induced* from bitexts. The difficulty lies in the fact that the mapping of words in bitexts is not known in advance.

What follows is a brief description of an algorithm for aligning words in bitexts and its adaptation to the task of determining recurrent sound correspondences.

5.1. *The word-to-word model of translational equivalence*

Algorithms for word alignment in bitexts aim at discovering word pairs that are mutual translations. Since words that are mutual translations tend to co-occur more frequently than other word pairs, a straightforward approach is to estimate the likelihood of translational equivalence by computing a similarity function based on a co-occurrence statistic, such as mutual information, Dice's coefficient, or the χ^2 test. The underlying assumption is that the association scores for different word pairs are independent of each other.

Melamed (2000) shows that the assumption of independence leads to invalid word associations, and proposes an algorithm for inducing models of translational equivalence that outperforms the models that are based solely on co-occurrence counts. His models employ the *one-to-one assumption*, which formalizes the observation that most words in bitexts are translated to a single word in the corresponding sentence. The algorithm, which is related to the expectation-maximization (EM) algorithm, iteratively re-estimates the *likelihood scores* which represent the probability that two word types are mutual translations. In the first step, the likelihood scores are initialized according to the G^2 statistic (Dunning, 1993), using only the co-occurrence information. Next, the likelihood scores are used to induce a set of one-to-one *links* between word tokens in the bitext. The links are determined by a greedy *competitive linking* algorithm, which proceeds to link pairs that have the highest likelihood scores. After the linking is completed, the link counts are used to re-estimate the likelihood scores, which in turn are applied to find a new set of links. The process is repeated until the translation model converges to the desired degree.

Melamed presents three translation-model estimation methods. Method A estimates the likelihood scores as the logarithm of the probability of jointly generating a pair of words. In Method B, an explicit noise model with auxiliary parameters is constructed in order to improve the estimation of likelihood scores. In Method C, bitext tokens are divided into classes, such as content words, function words, punctuation, etc., with the aim of producing more accurate translation models. The auxiliary parameters are estimated separately for each class.

5.2. *Adaptation*

Thanks to its generality and symmetry, Melamed's parameter estimation process can be adapted to the problem of determining correspondences. The main idea is to induce a model of sound correspondence in a bilingual word list, in the same way one induces a model of translational equivalence among words in a parallel corpus. After the model has converged, phoneme pairs with the highest likelihood scores represent the most likely correspondences.

While there are strong similarities between the task of estimating translational equivalence of words and the task of determining recurrent correspondences of sounds, a number of important modifications to Melamed's original algorithm are necessary in order to make it applicable to the latter task. The modifications include the method of finding a good alignment, the handling of null links, and the method of computing the alignment score.

The most important modification of the original algorithm concerns the method of aligning the segments in two corresponding strings. In sentence translation, the links frequently cross and it is not unusual for two words in different parts of sentences to correspond. On the other hand, the processes that lead to link intersection in diachronic phonology, such as *metathesis*, are sporadic. By imposing the no-crossing-links constraint on alignments, a dramatic reduction of the search space is achieved, and the approximate *competitive linking algorithm* of Melamed can be replaced with a variant of the algorithm for computing the optimal alignment of two strings (Wagner and Fischer, 1974).

Null links in statistical machine translation are induced for words on one side of the bitext that have no clear counterparts on the other side of the bitext. Melamed's algorithm explicitly calculates the likelihood scores of null links for every word type occurring in a bitext. In diachronic phonology, the pronunciation of any particular phoneme often changes over time, but it is rare for a phoneme to disappear without a trace across the entire lexicon. Therefore, insertion and deletion are modeled by employing a constant penalty for unlinked segments.

The alignment score is computed by summing the number of induced links and applying a small constant penalty for each unlinked segment, with the exception of the segments beyond the rightmost link. The exception reflects the relative instability of word endings in the course of linguistic evolution. In order to avoid inducing links that are unlikely to represent recurrent sound correspondences, only the phoneme pairs with likelihood scores above a set threshold are linked. (The threshold is established on a separate development set.) All correspondences above the threshold are considered to be equally valid. If more than one best alignment exists, links are assigned a weight averaged over the entire set of best alignments; for example, a link present in only one of two competing alignments receives the weight of 0.5. Finally, the score is normalized by dividing it by the average of the lengths of the two words.

The three methods described in Section 5.1 are adapted to the new task with minor modifications. In Method C, phonemes are divided into two classes: non-syllabic (consonants and glides) and syllabic (vowels); links between phonemes belonging to different classes are not induced. In addition, we propose a new Method D, which differs from Method C by excluding all links that include vowel phonemes. Method D emphasizes consonant correspondences, which are known to be much more stable.

	Proto-Algonquian	Cree	Fox	Menomini	Ojibwa
‘foam’	*pī**ʔt**ēw-	pī**st**ēw	—	pē**ʔt**ɛw	pī**tt**ē
‘grain’	*ke**ʔt**wikāni	ki**st**ikān	ke**ht**ikāni	—	—
‘tree’	*me**ʔt**ekwa	mi**st**ik	me**ht**ekwa	mɛ**ʔt**ek	mi**tt**ikw
‘sinew’	*a**ʔt**ehsi	a**st**is	—	a**ʔt**ɛh	—

Table 3. *Examples of complex recurrent sound correspondences between related languages*

6. Determination of complex recurrent sound correspondences

The algorithm described in the previous section can only discover correspondences between single phonemes. This limitation, which is directly inherited from Melamed's original algorithm, may prevent the algorithm from detecting complex (many-to-many) correspondences, such as the ones in Table 3. A quite similar problem exists also in the statistical machine translation. *Non-compositional compounds* (NCCs) are word sequences, such as “high school,” whose meaning cannot be synthesized from the meaning of its components. Since many NCCs are not translated word-for-word, their detection is essential in most NLP applications. In diachronic phonology, NCCs offer a limited method of capturing context-dependent correspondences.

6.1. *Discovery of non-compositional compounds in bitexts*

As a way of relaxing the *one-to-one* restriction, Melamed (1997) proposes an elegant algorithm for discovering NCCs in bitexts. His information-theoretic approach is based on the observation that treating NCCs as a single unit rather than as a sequence of independent words increases the predictive power of statistical translation models. Therefore, it is possible to establish whether a particular word sequence should be considered a NCC by comparing two translation models that differ only in their treatment of that word sequence. For the objective function that measures the predictive power of a translation model Melamed selects *mutual information*. Melamed's approach to the identification of NCCs is to induce a *trial translation model* that involves a candidate NCC and compare the model's total mutual information with that of a *base translation model*. The NCC is considered valid only if there is an increase of the mutual information in the trial model. In order to make this procedure more efficient, Melamed proposes inducing the translation model for many candidate NCCs at the same time. A complex gain-estimation method is used to guess whether a candidate NCC is useful *before* inducing a translation model that involves this NCC.

Given parallel texts E and F, the algorithm iteratively augments the list of NCCs. The iteration starts by inducing a base translation model between E and F. All continuous bigrams which are estimated to increase mutual information of the translation model are placed on a sorted list of candidate NCCs, but for each word token, only the most promising NCC that contains it is allowed to remain on the list. Next, a trial translation model is induced between E' and F, where E' is obtained from E by fusing each candidate NCC into a single token. If the net change in mutual information gain contributed by a candidate NCC is positive (i.e., greater than zero), all occurrences of that NCC in E are permanently fused; otherwise the candidate NCC is placed on a stop-list. The entire iteration is repeated until it reaches an application-dependent stopping condition.

6.2. *Adaptation*

The NCC algorithm is adapted with one major change. After inducing a trial translation model between E' and F, the original algorithm accepts all candidate NCCs that contribute a positive net change in mutual information gain. For the detection of phoneme NCCs, the modification is to accept all candidate NCCs that result in a correspondence that has a likelihood score above the minimum-strength threshold t, which is an adjustable parameter. We found that the strength of an induced correspondence better reflects the importance of a phoneme cluster than the mutual information gain criterion.

When the NCC approach is applied, the computation of the similarity score is slightly modified. Segments that represent valid NCCs are fused into single segments before the optimal alignment is established. The contribution of a valid correspondence is weighted by the averaged length of the correspondence. For example, a correspondence that links three segments on one side with two segments on the other side is given the weight of 2.5. As before, the score is normalized by dividing it by the average of the lengths of the two words. Therefore, the score for two words in which all segments participate in links is still guaranteed to be 1.0.

The algorithm terminates if two subsequent iterations fail to produce any candidate NCCs, or after a specific number of model-inducing iterations. In the experiments described in Sections 9 and 10, the maximum number of iterations of the algorithm was set to 12.

7. Semantic similarity

Since cognates originate from a single proto-form, many of them have either identical or similar meanings. Dictionaries and vocabulary lists usually define the meanings of the words in the form of *glosses* (cf. Table 2). Therefore, semantic similarity of two words can often be detected by comparing their respective glosses.

We investigated three increasingly sophisticated semantic similarity detection methods: Method G considers gloss identity only, Method K adds keyword-matching, and Method W employs also WordNet relations. In the absence of a WordNet-type resource, Method K can still be used provided that a part-of-speech tagger is available

Gloss A	Gloss B	Reason of mismatch
‘sweet grass’	‘sweetgrass’	spelling variants
‘ash’	‘ashes’	morphological variants
‘a mark’	‘mark’	redundant determiner
‘small stone’	‘stone’	adjectival modifier
‘goose’	‘snow goose’	nominal modifier
‘stone’	‘stone of peach’	complement
‘island’	‘island in a river’	adjunct
‘grave’	‘tomb’	synonymy
‘fowl’	‘turkey’	minor semantic shift
‘broth’	‘grease’	radical semantic shift

Table 4. *Examples of pairs of glosses that indicate semantically related words*

for the glossing meta-language. Otherwise, Method G is a fallback option. The three methods are discussed in detail in the following sections.

7.1. *Gloss identity*

The simplest method to detect semantic similarity is to check if the lexemes have one or more glosses in common. For example, Cree *kottāčīwin*‘terror, fear’ and Ojibwa *kostāčīwin*‘fear, alarm’ are correctly associated by this method. However, in many cases, the similarity of semantically related glosses is not recognized by this method. Table 4 contains some specific examples.

A large number of semantically related glosses are identified by employing relatively simple methods. Morphological variants are associated by means of lemmatization. Determiners, possessive pronouns, and very common modifiers, such as *certain, kind of, his, big, female,* etc. are placed on a stop-list and removed in the preprocessing stage. On the other hand, attempts at re-analyzing compounds written as a single word, as in the first example in Table 4, produce numerous false analyses (*thou-sand*, etc.).

7.2. *Keyword matching*

Many glosses contain phrases that include various modifiers, complements, and adjuncts, which often correspond to a common phenomenon of minor semantic shifts. Note that simple co-occurrence of words in glosses is not necessarily an indication of semantic similarity, e.g., ‘snow goose’ and ‘snow boots.’ One solution is to determine *keywords* — words that are likely to carry the meaning of a gloss. Pairs of glosses that contain matching keywords tend to be semantically related.

‘ **string**^{NN} for^{IN} stretching^{VBG} hide^{NN} ’
‘ upright^{JJ} **ornament**^{NN} worn^{VBN} on^{IN} head^{NN} ’
‘ yellow^{JJ} **feather**^{NN} with^{IN} black^{JJ} tip^{NN} ’
‘ **sorcerer**^{NN} who^{WP} has^{VBZ} a serpent^{NN} ’
‘ **clot**^{NN} of^{IN} **blood**^{NN} ’
‘ **flint**^{NN} , detonating^{VBG} **cap**^{NN} on^{IN} cartridge^{NN} ’
‘ **snow**^{NN} **dart**^{NN} , **ice**^{NN} throwing^{VBG} stick^{VB} ’
‘ **sign**^{NN} which^{WDT} points^{NNS} the way^{NN} ’
‘ a **portage**^{NN} , setting^{VBG} ashore^{RB} ’
‘ little **story**^{NN} that^{WDT} is^{VBZ} sometimes^{RB} told^{VBN} ’
‘ **mysterious**^{JJ} , haunted^{VBN} **person**^{NN} or^{CC} **place**^{NN} ’

Table 5. *Examples of automatically tagged glosses with keywords marked*

The following simple heuristic can be used to select noun keywords in glosses on the basis of the output of a part-of-speech tagger (Brill, 1995). Since the tagger operates on sentences rather than on phrases, all glosses are initially prepended with the string ‘It is a’ (e.g., ‘clot of blood’ is converted into ‘it is a clot of blood’). The string is removed after the tagging process is completed. Only words with the NN tag (nouns) are considered as possible keywords, except when a gloss contains a single word, in which case the word is taken to be the keyword regardless of the tag. The gloss is scanned from left to right, and all nouns are marked as keywords until a wh-word or a preposition other than ‘of’ is encountered.

Table 5 contains examples of keyword selection in action. The keywords identified by the heuristic are shown in boldface. Penn Treebank part-of-speech tags are provided for each word: prepositions are tagged as IN, while wh-words have tags that start with W. Stop-words are shown in a sans serif font.

It is evident from the handful of examples that the keyword scheme is far from perfect. Because of the limited accuracy of the part-of-speech tagger, some words are mistagged to begin with (e.g., ‘stick’). A comma separating two adjectives, as in ‘mysterious, haunted’ is indistinguishable from a comma separating two alternative glosses, so ‘mysterious’ is erroneously assumed to be an independent gloss. Nevertheless, the heuristic seems to pick most of the relevant nouns with reasonable precision.

We investigated two other methods of selecting keywords. The first method, which takes advantage of the simple syntax of glosses that can be modeled with a finite-state grammar, uses a finite-state transducer to select keywords from among the part-of-speech tags. The second method is to parse the glosses with a dependency parser and extract the syntactical heads of the phrases. In experiments, the first method produced almost exactly the same results as the heuristic described above, while the second method led to slightly lower overall accuracy.

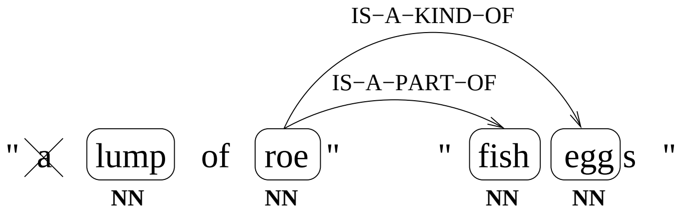

Figure 2. *An example of a partial semantic match between glosses*

7.3. *WordNet relations*

In order to identify semantically related glosses that contain no matching substrings whatsoever, it is necessary to refer to some lexical resource. Lowe and Mazaudon (1994, footnote 13, page 406) suggest using WordNet (Fellbaum, 1998) for the detection of semantic relationships. WordNet's noun hierarchy is particularly well suited not only for detecting synonyms but also for associating lexemes that have undergone minor semantic changes. Trask (1996) lists several types of semantic change, including the following:

– **generalization** (broadening): 'partridge' → 'bird';

– **specialization** (narrowing): 'berry' → 'raspberry';

– **melioration** (developing a more favorable sense): 'woman' → 'queen';

– **pejoration** (developing a less favorable sense): 'farm-worker' → 'villain';

– **metaphor** (extending the literal meaning): 'steersman' → 'governor';

– **metonymy** (using an attribute of an entity to denote the entity itself): 'crown' → 'king';

– **synecdoche** (using a part to denote a whole, or vice-versa): 'hand' → 'sailor'.

Certain types of semantic change have direct parallels among WordNet's lexical relations. *Generalization* can be seen as moving up the IS-A hierarchy along a hypernymy link, while *specialization* is moving in the opposite direction, along a hyponymy link. *Synecdoche* can be interpreted as a movement along a meronymy/holonymy link. However, other types of semantic change, such as metonymy, melioration/pejoration, and metaphor, have no direct analogues in WordNet.

One possible approach to the calculation of a WordNet-based semantic similarity score is to define it as a function of the length of the shortest path between synsets, measured in the number of IS-A links, e.g., normalized path length of Leacock and Chodorow (1998). However, our preliminary experiments indicated that the effect of considering paths longer than one link was negligible.

A simpler solution is to consider only synsets directly linked by a relationship link, and estimate the semantic similarity on the basis of the type of link and whether

Word	Synonyms	Hypernyms	Meronyms
lump	ball, clod, glob, clump, chunk, swelling, klutz, puffiness, lout, clod, goon, stumblebum, oaf, lubber, lummox, gawk, hunk	agglomeration, piece, part, symptom, clumsy person	—
roe	hard roe	spawn, **egg**, seafood	**fish**
fish	chump, fool, gull, mark, patsy, fall guy, sucker, shlemiel, soft touch, mug, go fish	foodstuff, food product, victim, card game, cards, dupe, aquatic vertebrate	pisces, school, shoal
egg	testis, gonad, testicle, ball, ballock, bollock, nut	endocrine gland, ductless gland, ovum, egg cell, foodstuff, food product	male genitalia, family jewels

Table 6. *Lists of semantically related words extracted from WordNet*

it applies to the entire gloss or just a keyword. Four lexical relations (*identity*, *synonymy*, *hypernymy*, and *meronymy*) and two focus levels (*gloss* and *keyword*) yield eight semantic similarity features.

In our implementation, the lemmatization process is carried out by *QueryData*, a *Perl* interface to WordNet developed by Jason Rennie. Glosses that exceed 30 characters are truncated. A list of synonyms, hypernyms, and meronyms is then generated for each gloss and keyword. Words are considered to be related if there is a relationship link between any of their senses.

The entire process of detecting semantic similarity between vocabulary entries can be traced using an example involving Cree *wāhkwa* 'a lump of roe' and Ojibwa *wākk* 'fish eggs' (Figure 2). After the preprocessing removes the determiner a from the first gloss, the glosses are tagged with a part-of-speech tagger, and the following four nouns are identified as keywords: lump, roe, fish, eggs. The lemmatization removes the plural ending -s from eggs. Neither of the complete glosses exists in WordNet, but each of the keywords is represented by several senses. The WordNet sense lists for the keywords are shown in Table 6. In the end, two semantic similarity features are detected: roe is a kind of egg (*keyword hypernymy*), and roe is a part of fish (*keyword meronymy*).

The use of WordNet for semantic similarity detection is possible only if English is the glossing metalanguage. If the available vocabularies are glossed in other languages, one possible solution is to translate the glosses into English, which, however, may increase their ambiguity. A better solution would be to employ a multilingual lexical resource, such as EuroWordNet (Vossen, 1998), which is modeled on the original WordNet. In general, WordNet could be replaced by another machine-readable dictionary or thesaurus, or even semantic similarity clusters automatically extracted from a corpus.

8. Combining various types of evidence

For the purpose of identifying cognates, we consider three different types of evidence. The phonetic and the correspondence-based approaches produce continuous scores in the [0,1] range. The semantic approach supplies a vector of eight binary semantic features. The task is to combine all three types of evidence into a single numerical score, which will be used for ranking potential cognate pairs. Our initial approach presented in (Kondrak, 2001) was to manually assign numerical scores to the eight semantic features and then linearly combine the resulting semantic similarity score with the other two continuous scores. In (Kondrak, 2004), we presented a Naive Bayes approach, in which all features are assumed to be independent. Below, we describe the method for converting the similarity scores into probabilities by means of the Beta distributions, and the treatment of the semantic features. A detailed description of the method for computing the overall similarity score for a pair of words can be found in (Kondrak, 2004).

First, we use the Beta distribution to convert each of the two continuous scores into a probability that two words are related. The Beta distribution is defined over the domain $[0, 1]$. Figure 3 shows the fit between the distributions of scores between word pairs in our development set and the corresponding Beta distributions. The development set contains a large number of word pairs generated by taking a Cartesian product of Cree and Ojibwa vocabulary lists. The scores are clustered within the 0.04 intervals. The parameters of the corresponding Beta distributions were calculated from the mean and variance of the scores. The Beta distributions fit the distributions of scores quite well. The fit with the phonetic scores of unrelated words is particularly good. For cognate words, the fit is somewhat less tight, which is not surprising considering that the number of cognate pairs is several magnitudes times smaller than the number of unrelated pairs. In the case of correspondence-based scores, the Beta distribution fails to account for a number of cognate pairs that are completely covered by correspondences (score = 1.0). This problem is likely to be less acute for language pairs that are not as closely related as Cree and Ojibwa because the number of such word pairs is expected to be much smaller. The parameter values established on one language pair can be used for language pairs that have no cognation information, which is demonstrated in Section 10.

With regards to the semantic features, the main difficulty lies with their interdependence. For instance, the fact that a pair of glosses are synonymous increases the probability that their respective keywords are also synonymous. One solution is to define a *subsumption hierarchy* of the features, and disregard a feature if a feature that dominates it in the hierarchy is present. For example, it seems rather obvious that if a keyword identity is detected, there is no further advantage to considering keyword hypernymy or meronymy. We investigated several partial and linear order-

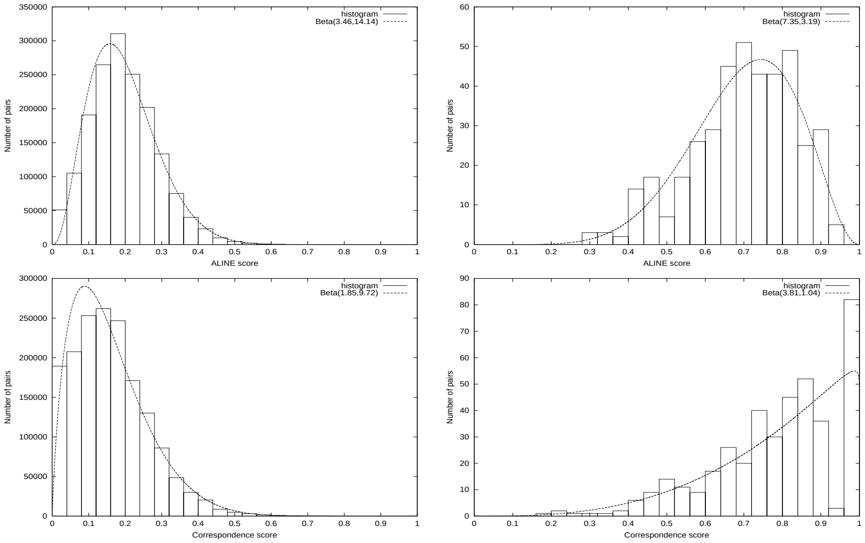

Figure 3. *Distribution of the phonetic scores (top) and the correspondence-based scores (bottom) for the unrelated (left) and the cognate (right) word pairs, and the corresponding Beta distributions*

ings of features and concluded that a straightforward linear ordering is hard to surpass (Kondrak, 2004). The following linear ordering was established empirically on the development set: *gloss identity* > *gloss synonymy* > *keyword identity* > *gloss hypernymy* > *keyword synonymy* > *keyword hypernymy* > *gloss meronymy* > *keyword meronymy*. We briefly discuss the effect of other feature orderings on the overall accuracy in Section 10.

The advantage of the initial approach is that it does not require annotated training data. In the Naive Bayes approach, a number of parameters must be established on a separate training set. However, the values of the parameters are set automatically rather than manually.

9. Experiments on bilingual word lists

In this and the following two sections, we describe experiments aimed at identifying correspondences and cognates in Indo-European bilingual word lists (Section 9), Algonquian vocabulary lists (Section 10), and Totonac-Tepehua vocabulary lists (Section 11). The difference between a bilingual word list and a pair of vocabulary lists can be seen by comparing Tables 1 and 2. We test our three computer programs: ALINE (Section 4.2), which computes phonetic similarity between a pair of words, CORDI, which detects correspondences in bilingual word lists, and COGIT, which identifies cognates shared by vocabulary lists on the basis of various types of evidence. In the Sections 9 and Section 10, we compare the results of our programs to “gold standards” established by historical linguists. In Section 11, we apply them to a pair of languages from a relatively little-studied family, which is yet to be thoroughly analyzed.

A bilingual word list is a collection of word pairs from two languages where the corresponding words have the same, well-defined meaning. One of the most widely used set of meanings is the list of 200 basic words that are relatively resistant to lexical replacement and exist in most of the world’s languages (Swadesh, 1952). It includes body parts (*hand, neck, nose*), actions (*breathe, play, spit*), animals (*bird, dog, snake*), etc. The Swadesh word lists have been compiled for many of the world’s languages. Nevertheless, the methods we propose are equally applicable to other sets of meanings.

9.1. *Data sets*

The development set consisted of six 200-word lists representing Italian, Polish, Romanian, Russian, Serbo-Croatian, and Spanish, adapted from the Comparative Indo-European Data Corpus (Dyen *et al.*, 1992). The cognation judgments which served as our gold standard were originally made by Isidore Dyen. We manually transcribed the lists from a restricted orthographic representation into an IPA-like phonetic notation.

The test set consisted of five 200-word lists representing English, German, French, Latin, and Albanian, compiled by Kessler (2001) As the lists contain rich phonetic and morphological information, the stemmed forms were automatically converted from the XML format with virtually no extra processing. The gold standard included only the cognation judgments that were annotated as certain by Kessler.

The language pairs in the test set, except the English-German and the French-Latin pairs, are quite challenging for a cognate identification program. In many cases, the gold-standard cognate judgments distill the findings of decades of linguistic research. In fact, for some of those pairs, Kessler finds it difficult to show by statistical techniques that the surface regularities are unlikely to be due to chance. Nevertheless, in order to avoid making subjective choices, our programs were evaluated on all ten language pairs.

9.2. *Determination of correspondences in word lists*

We implemented the methods for the identification of correspondences described in Sections 5 and 6 as a C++ program, named CORDI. The initial experiments indicated that CORDI has little difficulty in determining correspondences given a set of *cognate* pairs from a pair of related languages. However, the very existence of a reliable set of cognate pairs implies that the relationship between the languages in question has already been thoroughly investigated, and that the sound correspondences are known. A more realistic input for the program is a bilingual word list that con-

corr	*cooc*	*links*	*cogn*	*score*	*corr*	*cooc*	*links*	*cogn*	*score*
r:r	26	24	9	158.7	**l:l**	14	9	3	49.7
n:n	24	23	13	154.2	**h:k**	7	7	4	47.6
t:d	18	18	7	122.4	l:f	9	7	1	43.0
k:k	12	11	0	72.5	j:g	6	6	1	40.8
s:s	11	10	8	65.7	j:k	10	7	4	40.7
f:p	9	9	7	61.2	m:w	7	6	1	38.5
m:m	10	9	3	58.9	d:b	5	5	1	34.0
d:t	10	8	3	49.8	**θ:t**	6	5	3	31.7

Table 7. *English-Latin correspondences discovered by CORDI in a bilingual word list. The historically valid correspondences are shown in bold*

tains both cognate pairs and unrelated word pairs. Determining correspondences in a bilingual word list is clearly a more challenging task than extracting them from a list of reliable cognates because the non-cognate pairs introduce noise into the data.

In order to test the ability of our system to identify correspondences in noisy data, we applied Method D to the English-Latin bilingual word list. Only 24.5% of word pairs in the list are actually cognate; the remaining 75.5% of the pairs are unrelated. A model for a 200-pair list usually converges after 3–5 iterations, which takes only a few seconds on a standard PC.

Table 7 shows the correspondences determined by CORDI sorted by their likelihood scores. In total, nine of the sixteen correspondences are valid (shown in bold), including eight among the top ten. In contrast, only five of the top sixteen phoneme matchings picked up by the χ^2 statistic are valid correspondences. According to Watkins (2000), there are about a dozen additional consonant correspondences between English and Latin, but only one of those (**w:w**) appears more than once among the cognate pairs in our word list.

Unlike statistical approaches, CORDI produces explicit alignments, which makes it possible to trace the correspondences to individual word pairs. In most cases, the number of times the program posits a correspondence (the *links* column) is very close to the number of times the two phonemes co-occur in the data (the *cooc* column). The number of correspondence links that occur in *cognate* pairs is given in the *cogn* column. The correct correspondence links must satisfy two conditions: (1) they must occur in cognate pairs and (2) they must represent historically valid correspondences. We manually verified that *all* correspondence links posited by CORDI in the English-Latin word list that satisfy the above two conditions are indeed correct. However, some of the links posited in cognate pairs represent invalid correspondences. For example, CORDI posits two links between the cognate pair [najt] and [nokt] ‘night’: **n:n** and **j:k**, of which only the former is correct. On the other hand, both the **r:r** and

Phonetic-based methods		Correspondence-based methods		Combined methods	
PREFIX	.544	COGNATE	.516	Method D + NCC	.619
DICE	.467	Method A	.515	ALINE + Method D	**.681**
LCSR	.561	Method B	.565	ALINE + Method D + NCC	.677
JAKARTA	.513	Method C	.580		
ALINE	.628	Method D	.629		

Table 8. *The average cognate identification precision obtained by various methods on the test set composed of Indo-European bilingual word lists*

k:k links posited between the unrelated [bɑrk] and [kortik] 'bark' are incorrect, even though the **r:r** correspondence is valid in general.

9.3. *Identification of cognates in word pairs*

Because of the lack of a readily available gold standard, the quality of correspondences produced by CORDI is difficult to validate, quantify, and compare with the results of alternative approaches. We are interested in a broad evaluation involving a number of different language pairs using an independently developed gold standard. The Comparative Indo-European Data Corpus specifies which word pairs are cognate, but does not list recurrent correspondences. Simply counting the number of correspondences in related word pairs is not satisfactory because, as we have demonstrated in the previous section, the fact that a correspondence link is posited in a cognate pair does not necessarily imply that the link is valid. However, since the likelihood of cognation of a pair of words increases with the number of correspondences that they contain, it is possible to evaluate the correspondences indirectly by using them to identify cognates.

The evaluation method adopted for measuring the effectiveness of cognate identification programs is the *11-point interpolated average precision*. A cognate identification program is applied to a bilingual word list, and produces as output a list of the candidate word pairs sorted by their scores. Typically, true cognates are very frequent near the top of the list, and become less frequent towards the bottom. The threshold cut-off value may depend on the intended application, the degree of relatedness between languages, and the particular method used. Rather than reporting precision and recall values for an arbitrarily selected score threshold, precision is computed for the recall levels of 0%, 10%, 20%, . . . , 100%, and then averaged to yield a single number. At recall level x, precision is computed at the point of the ranked list where the proportion of identified true cognate pairs reaches x. A perfect ordering of all cognate pairs before all non-cognate pairs translates into a 1.0 precision. The expected precision of a random ordering of word pairs is close to the proportion of cognate pairs in the list.

Table 8 compares the average precision achieved on the test set by various methods proposed in this article and by two programs mentioned in Section 3, COGNATE and JAKARTA. Statistical significance was computed with the t-test for independent samples. ALINE significantly outperforms other phonetic and orthographic methods, including JAKARTA. Among the correspondence-based approaches, COGNATE is about as accurate as Method A, but both are outperformed by the methods that employ an explicit noise model. Method D, which considers only consonants, achieves the highest precision. In spite of its extra complexity, Method C is not significantly better than Method B. The top phonetic method (ALINE) and the top correspondence-based method (Method D) obtain similar average cognate identification precision.

The combination of the top phonetic and correspondence-based methods achieves the highest precision among all the methods that were tested. The top phonetic and correspondence-based methods were combined using the method described in Section 8. The adjustable parameters were derived from the Italian-Polish word list, with both types of evidence weighted equally. The Italian-Polish language pair was chosen because it produced the best overall results on the development set, but the relative differences in average precision with different training sets did not exceed 1%. We also tried other combinations of methods (not shown in Table 8), but they perform worse.

Somewhat surprisingly, the incorporation of complex correspondences (the NCC approach) has a slightly negative effect on the results. A close examination of the results indicates that few useful complex correspondences were identified by the NCC algorithm in the 200-word Indo-European lists. This may be caused by the small overall number of cognate pairs (57 per language pair, on average), or simply by the paucity of recurrent complex correspondences.

Another surprising finding was that straightforward averaging of the phonetic and the correspondence-based scores produces results that are quite similar to the results obtained using the method described in Section 8. On the test set, the straightforward method achieves the average precision of 0.686 for ALINE combined with method D, and 0.685 for the same approach utilizing complex correspondences (NCC). However, the averaging of two different kinds of scores has little theoretical justification, and there is no guarantee that such an approach would work well on other language sets.

Figure 4 breaks down the results obtained by one of the top methods by language pair. It is evident that the identification precision is highly correlated with the proportion of cognate pairs in the list. The identification precision is very good for closely related languages, but falls as the noise-to-signal increases. In the extreme case, the Albanian-English list contains only 20 cognates versus 180 unrelated pairs. Still, six out of ten top-ranked pairs in that list are cognates.

In order to illustrate how the phonetic similarity and recurrent sound correspondences balance each other, we include a few examples from the English-Latin word list. An unrelated [de]/[diē] 'day' is ranked 16 among 200 pairs by phonetic similarity, but the lack of correspondences brings down the overall ranking to 80. On the

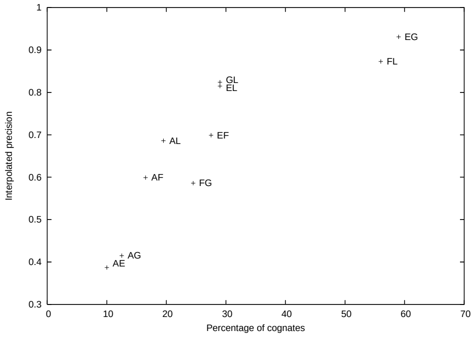

Figure 4. *The proportion of cognates and the 11-point interpolated average cognate identification precision obtained by the top method from Table 8 on all language pairs in the test set. The languages are: English (E), German (G), French(F), Latin(L), and Albanian(A)*

other hand, cognates [hu]/[kwo] (stem of *quis*) ‘who’ phonetically rank 124, but the presence of the **h:k** correspondence increases the overall rank to 62. Phonetic similarity also boosts the rankings of cognate pairs with less regular correspondences (e.g., [stɑr]/[stēlla] ‘star’). The only cognate pair with both partial scores equal to zero is [ɛg]/[ōwo] ‘egg’. As a result, among the top 40 pairs in the list, there are only four pairs that are marked as unrelated in our gold standard. Three of them, [flo]/[flue] ‘flow’, [skræʧ]/[skabe] ‘scratch’, and [spɪt]/[spue] ‘spit’, are due to high phonetic similarity, while the fourth one, [dərti]/[sordido] ‘dirty’, contains two specious correspondences (**r:r** and **t:d**).

An advantage of our algorithm is its capability of linking phonemes in any word position. The approaches that rely on the syllabic structure of words (Ringe, 1992) or the character position within a word (Tiedemann, 1999) tend to produce rigid alignments, which are unable to handle phenomena such as epenthesis (insertion of a vowel between consonants) or syncope (loss of a vowel between consonants). In the English-Latin list, alignments that do not follow the syllabic structure include [fʊl]/[plēno] ‘full’ and [ni]/[genū] ‘knee.’

It is tempting to apply the program to languages which are presumed to be unrelated. However, any bilingual word list of considerable size is likely to contain a number of accidental regularities. For example, Hawaiian [l] and Turkish [k] co-occur in 21 out of 200 word pairs in the corresponding list, which also includes such phonetically similar word pairs as [hele]/[gel] ‘to come’ and [omo]/[em] ‘to suck’. Although most of the correspondences identified in lists of unrelated words are clearly phonetically implausible, and the top similarity scores are somewhat lower than for lists representing remotely-related languages, no attempt was made to interpret the output in terms of the probability of a genetic relation between languages. We believe that statistical tests of the kind proposed by Ringe (1998) and Kessler (2001) are better suited for this purpose.

10. Experiments on vocabulary lists

In this section, we describe experiments aimed at identifying correspondences and cognates between pairs of vocabulary lists. A vocabulary list is a list of lexemes from a single language accompanied by glosses in another language that explain their meaning (cf. Table 2). Glosses may be either single words or complex phrases. One method of obtaining a vocabulary list is by automatically scanning a traditional bilingual dictionary.

10.1. *Data sets*

The Algonquian data set consists of two parts that complement each other: the etymological dictionary (Hewson, 1993) and the vocabulary lists from which the dictionary was produced. The dictionary, which served as a source of the cognation information, contains 4,068 cognate sets, including 853 marked as nouns. The lexemes are already in a phonemic transcription, so no elaborate grapheme-to-phoneme conversion was necessary. The vocabulary lists represent the four principal Algonquian languages, Fox, Menomini, Cree, Ojibwa, and contain over 27,000 entries in total, including almost 5,000 noun entries. The development set consisted of the Cree-Ojibwa language pair, while the remaining five pairs served as the test set. In contrast with the dictionary, the vocabulary lists contain many errors, inconsistencies, duplicates, and lacunae. Only limited, automatic validation of the data had been performed, which removed entries that were clearly duplicate or explicitly marked as doubtful. Although manual correction of erroneous individual entries and cognation judgments would undoubtedly improve the accuracy, a noisy data set provides a more trustworthy test for a system designed to help solve real linguistic problems.

10.2. *Identification of correspondences*

The first step towards identifying recurrent correspondences between two vocabulary lists is to automatically construct a bilingual word list that contains a sufficiently high proportion of cognates. The method that we adopted was to extract, from a Cartesian product of all entries, all pairs of noun lexemes that had at least one gloss in common. The resulting bilingual word lists were composed of both cognate and unrelated pairs: the development set (Cree-Ojibwa) contained 732 pairs, including 242 (33.1%) cognate pairs, while the test set (Fox-Menomini) contained 397 word pairs, including only 79 (19.9%) cognate pairs.

Since the vowel correspondences in Algonquian are rather inconsistent, following Hewson (1974), we decided to concentrate on consonants and consonant clusters.

One type of evidence		Two types of evidence		Three types of evidence	
PH	.430	PH+SC	.472	PH+CC+SEM(G)	.633
SC	.448	PH+CC	.513	PH+CC+SEM(K)	.649
CC	.473	PH+SEM(W)	.631	PH+CC+SEM(W)	**.660**
SEM(W)	.227	CC+SEM(W)	.625	PH+SC+SEM(W)	.652

Table 9. *The average cognate identification precision obtained by various methods on the test set composed of Algonquian vocabulary lists. The types of evidence are: phonetic similarity (PH), simple correspondences (SC), complex correspondences (CC), and semantic similarity (SEM). The semantic methods are: gloss identity only (G), gloss and keyword identity (K), and the WordNet-based method (W)*

Method C was selected for the evaluation on the basis of the experiments involving the development set. On the Fox-Menomini data, the algorithm terminated after 12 iterations, which took several minutes on a Sparc workstation. (Each iteration involves inducing anew both the base and the trial translation models.)

Table 10 compares the sets of correspondences identified in the Fox-Menomini data by JAKARTA (Oakes, 2000), Method C (simple correspondences), and Method C augmented with the NCC approach (complex correspondences). The results were evaluated against the set of 31 correspondences enumerated by Bloomfield (1946), which contains 1:1, 1:2, and 2:2 consonant correspondences. Bloomfield's correspondences are shown in boldface in Table 10.

Both JAKARTA and Method C achieve high precision, but low recall. They identify 7 and 9 of the valid correspondences, respectively. Method C by itself can discover only simple, 1:1 correspondences. JAKARTA is capable in principle of identifying complex correspondences, but the only one it posits is incorrect. In contrast, Method C augmented with the NCC approach identifies 23 correspondences, 20 of which are correct.

In order to determine why the remaining 11 valid correspondences were not identified by CORDI, we manually analyzed the 79 cognate pairs included in the input word list. We established that the correspondences **š:hk** and **p:hp** occur twice in the input, **hč:ʔč** occurs once, and the remaining seven complex correspondences do not occur at all. The **h:ʔ** correspondence occurs in the list only within the clusters **hč:ʔč** and **ht:ʔt**. Since, by definition, recurrent correspondences are those that occur at least twice, both precision *and* recall obtained by CORDI on the test set were close to 90%.

10.3. *Identification of cognates in vocabulary lists*

A pair of vocabulary lists may contain many hundreds of cognates, but since words are not paired by their meanings as in a bilingual word list, finding them is more

Method	Correspondences
JAKARTA	**n:n k:k m:m p:p h:h s:s t:t** h:hs
Method C	**k:k n:n h:h m:m p:p š:s s:s t:t č:č** s:ʔ
Method C + NCC	**n:n k:k m:m h:h p:p hk:hk š:s t:t č:č s:s s:hs s:ʔs šk:sk** p:č hk:t **ht:ʔt hp:hp s:ʔn š:ʔs** t:sk **t:ht č:hč s:hn**

Table 10. *The Fox-Menomini consonantal correspondences determined by various methods. The historically valid correspondences are shown in bold*

difficult. To take the Menomini-Ojibwa pair as an example, the task of the system was to identify 259 cognate-pairs from among 1540 × 1023 possible lexeme-pairs, which means that there were about 6500 unrelated pairs for each cognate pair. On the other hand, it is possible to discover cognates whose meanings have shifted and which no longer are synonymous with one another. All three types of evidence are now available: semantic, phonetic, and correspondences. One of the goals of the experiment described in this section was to evaluate the contribution of individual types of evidence to the overall performance of the system.

COGIT, our implementation of the methods presented in this article, takes two vocabulary lists representing distinct languages as the input, and produces a list of vocabulary-entry pairs, sorted according to the estimated likelihood of cognation. COGIT can combine one, two, or all three types of evidence, and therefore subsumes both ALINE and CORDI. Table 9 compares the 11-point interpolated average precision achieved by various configurations on the test set. The table has three parts, which correspond to the number of sources of evidence. The values of tunable parameters were established during the development phase of the system, using the Cree-Ojibwa data.

The leftmost part of Table 9 includes methods that utilize only a single source of evidence. ALINE is selected as the representative of the phonetic methods, as it achieves significantly higher precision than other phonetic and orthographic approaches on word lists, especially on more remotely related language pairs. Method B represents the correspondence-based methods on the basis of its performance on the development set. This time, a pure correspondence-based approach outperforms the best phonetic method, especially when complex correspondences are utilized. Relying on gloss similarity alone (Method W) with no information from lexemes yields predictably low precision, because no continuous score is available to order candidate pairs within the semantic similarity classes.

The center part of Table 9 shows methods that combine two types of evidence. In all cases, there is an improvement over individual methods. The improvement is particularly dramatic when the evidence from glosses is combined with the evidence from lexemes. All methods that use the semantic information provided by the glosses perform substantially better than the methods that use only the information contained in lexemes.

	Test set						Dev. set
	FM	FC	FO	MC	MO	**Avg**	CO
Cognates	121	130	136	239	259	**0.015%**	408
Precision	.651	.698	.691	.618	.641	**.660**	.787

Table 11. *The number of cognates and the 11-point interpolated average cognate identification precision obtained by the top method from Table 9 on all language pairs involving Fox (F), Menomini (M), Cree(C), Ojibwa(O)*

The rightmost part of Table 9 shows the results when all three types of evidence are combined. Even when only gloss identity is considered (Method G), there is an impressive performance improvement in comparison to the purely lexeme-based methods. Adding keyword-matching (Method K) and WordNet relations (method W) brings an additional, albeit modest, improvements. The differences between alternative orderings of semantic features discussed in Section 8 are too small to warrant inclusion in the table. However, applying the features without any ordering is almost equivalent to using no semantics at all. Finally, we note that the advantage provided by complex correspondences all but disappears when all types of evidence are combined.

10.4. *The role of WordNet*

The reasons for the relatively small contribution of WordNet to the overall performance of the system can be attributed both to the properties of the test data and to the shortcomings of WordNet itself. Since the data for all Algonquian languages originates from a single project, it is quite homogeneous. As a result, many glosses from different vocabulary lists are identical within cognate sets, which limits the need for the application of WordNet lexical relations. In particular, 62% of all cognate pairs have identical glosses, and additional 10% have keywords in common. Method W is able to detect similarity in 28% of the remaining cognate pairs, which constitutes about 8% of all cognate pairs. The 20% of the glosses where no similarity is detected include glosses that, even after preprocessing that includes spell-checking and lemmatization, do not match any WordNet senses. This problem occurs for instance when a compound word is written as a single word (e.g., ‘sweetgrass’), or when a rare word is not included in WordNet at all (e.g., ‘spawner’).

The other source of errors are the semantic associations implied by WordNet’s lexical relations. Polysemous words are sometimes incorrectly associated on the basis of uncommon senses; for example, star and lead share a synset defined as ‘an actor who plays a principal role.’ On the other hand, some words that are semantically very similar, such as puppy and dog, happen to be far from each other in the WordNet hierarchy. Other incorrect associations include gun *is-a-part-of* airplane, (‘a pedal that controls the throttle valve’), and snare *is-a-kind-of* drum (‘a small drum with two heads and a snare stretched across the lower head’). In many cases, however, the associations provide a semantic link between cognates that cannot be correlated by simple string matching; for example, bachelor *is-a* unmarried man (synonymy), gooseberry *is-a-kind-of* currant (hyponymy), and mattress *is-a-part-of* bed (meronymy).

11. Experiments on previously unanalyzed data

Unlike the experiments described so far, which concerned well-studied language families, the final experiment involved languages whose mutual relationship is still being investigated. This was consistent with our goal of providing tools for the analysis of little-studied languages represented by word lists. On the other hand, the evaluation of the results was more difficult because of the paucity of confirmed sets of correspondences and cognates.

11.1. *The Totonac data set*

The final set of experiments was performed in the context of the Upper Necaxa Field Project. Upper Necaxa is a seriously endangered language spoken by a few thousand indigenous people in Puebla State, Mexico. The primary goal of the project is to document the language through the compilation of an extensive dictionary and other resources, which may aid revitalization efforts. One aim of the project is the investigation of the relationship between Upper Necaxa Totonac and the other languages of the Totonac-Tepehua language family, whose family tree is not yet well-understood.

The data for the experiment consisted of Spanish dictionaries of Upper Necaxa (Beck, 2001) and Sierra Totonac (Aschmann, 1983), which are available in electronic form. Both languages belong to the Totonac-Tepehua language family. After a preprocessing stage, which included automatic conversion of the lexemes from an orthographical into a phonetic notation, the nouns extracted from the dictionaries were analyzed by our system in order to identify recurrent correspondences and cognates. The Upper Necaxa list contained 2110 nouns, and the Sierra list contained 763 nouns.

11.2. *Identification of correspondences*

In the first experiment, CORDI, the correspondence-identification program, was applied to Upper Necaxa and Sierra. Simple correspondences were targeted, as complex correspondences do not seem to be very frequent among the Totonac languages. The input for CORDI was created by extracting all pairs of noun lexemes with identical glosses from the two dictionaries. The input list contained 865 word pairs, and was likely to contain more unrelated word pairs than actual cognates.

The correspondences were evaluated by David Beck, the principal investigator of the Upper Necaxa Field Project. Of the 24 correspondences posited by CORDI, 22 were judged as completely correct, while the remaining two (**ʧ:ʦ** and **tɬ:ʦ**). were judged as “plausible but surprising.” Since CORDI explicitly list the word pairs from which it extracts correspondences, they were available for a more detailed analysis. Of the five pairs containing **ʧ:ʦ**, one was judged as possibly cognate: Upper Necaxa [**ʧ**astun] and Sierra [aʔa**ʦ**astun] ‘*rincón*, *esquina*.’ Both word pairs containing **tɬ:ʦ** were judged as possibly cognate: [li**tɬ**an]/[li**ʦ**eχ] ‘*favor*,’ and [**tɬ**aq**tɬ**a]/[**ʦ**a**ʦ**a] ‘*elote*.’ Both unexpected correspondences were assessed as meriting further investigation.

11.3. *Identification of cognates*

In the second experiment, COGIT, the cognate identification program, was run on the vocabulary lists containing the Upper Necaxa and Sierra nouns. Because of a different glossing meta-language and the lack of Totonac training data, the overall approach was somewhat simpler than in the previous experiments. The keyword selection heuristic was simply to pick the first word of the gloss. More complex semantic relations were not considered. The three types of evidence were combined by a linear combination of scores, with gloss identity given double the weight of keyword identity.

A large list of the candidate word pairs with their glosses was sorted by the total similarity score and presented to the human judge. The judge was instructed to evaluate the pairs in order, starting from the top of the list, and to stop when the proportion of false positives became too high to justify further effort. The pairs were judged as true positives only if the word roots were cognate; sharing an affix was not deemed sufficient. Compound words were counted as cognates if any of the multiple roots were related; for example, both *snowstorm/storm* and *snowstorm/snow* would be acceptable. The rationale is that a person compiling an etymological dictionary would still want to know about such pairs whether or not they are eventually included as entries in the dictionary.

In total, 711 pairs were evaluated, of which 350 were classified as cognate, 351 as unrelated, and 10 as doubtful. 18 of the positive judgments were marked as loans from Spanish. In Figure 5, the boxes correspond to the precision values for the seven sets of 100 candidate pairs each, sorted by score; the curve represents the cumulative precision. As can be seen, almost all the pairs in the beginning of the file were cognates, but then the number of false positives increases steadily. In terms of semantic similarity, 30% of the evaluated pairs had at least one gloss in common, and a further 7% shared a keyword. Among the pairs judged as cognate, the respective percentages were 49% and 11%.

(Kondrak *et al.*, 2007) describes further experiments aimed at creating an etymological dictionary of the Totonac family of languages, which incorporated vocabulary lists representing three other related languages.

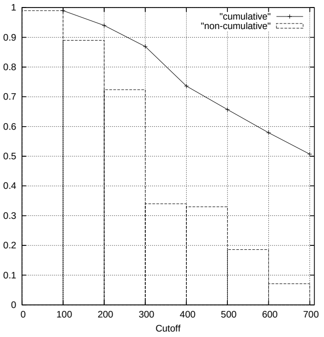

Figure 5. *Cognate identification precision on the Totonac test set*

12. Conclusion

We have presented novel methods for the identification of cognates and recurrent sound correspondences, applicable both to structured word lists and unstructured vocabulary lists. Our robust iterative approaches detect both simple and complex correspondences by exploiting the idea of relating correspondences between sounds to translational equivalences between words. Cognates are identified by combining three distinct types of evidence: recurrent sound correspondences, phonetic similarity of words, and semantic similarity of glosses,

We conducted thorough evaluation experiments involving three distinct language families. The results of our experiments demonstrate that our programs are more accurate than both comparable programs and purely statistical approaches, and perform well on noisy test sets in which the unrelated word pairs substantially outnumber the cognate pairs. Incorporating each of the three types of evidence clearly helps in cognate identification, regardless of the actual combination method. In particular, detecting semantic similarity seems to be crucial in unstructured vocabulary lists, but a sophisticated method based on WordNet offers little improvement over simple lexical matching. Even though complex correspondences in the Algonquian data are identified with excellent recall and precision, their incorporation does not result in finding more cognates.

The algorithms described here accomplish in mere minutes what could take many hours (perhaps years) of expert labor, given the large amounts of data that require processing. The final experiment was designed specifically to show that our programs can be applied in a realistic setting. We hope that the algorithms will assist historical linguists in their investigations of little-studied language families and in furnishing conclusive evidence for hitherto conjectural language groupings.

Acknowledgements

The research reported here was financially supported by the Natural Sciences and Engineering Research Council of Canada.

13. References

Adamson G. W., Boreham J., “The use of an association measure based on character structure to identify semantically related pairs of words and document titles”, *Information Storage and Retrieval*, vol. 10, p. 253-260, 1974.

Aschmann H. P., *Vocabulario totonaco de la Sierra*, Summer Institute of Linguistics, Mexico, 1983.

Baxter W. H., Ramer A. M., “Beyond lumping and splitting: probabilistic issues in historical linguistics”, *in* C. Renfrew, A. McMahon, L. Trask (eds), *Time Depth in Historical Linguistics*, p. 167-188, 2000.

Beck D., *Primer vocabulario práctico del idioma totonaco del Río Necaxa*, University of Alberta, 2001.

Bloomfield L., “Algonquian”, *in* H. Hoijer (ed.), *Linguistic Structures of Native America*, Viking, p. 85-129, 1946.

Brill E., “Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging”, *Computational Linguistics*, vol. 21, n° 4, p. 543-566, 1995.

Burton-Hunter S. K., “Romance Etymology: a computerized model”, *Computers and the Humanities*, vol. 10, p. 217-220, 1976.

Covington M. A., “An Algorithm to Align Words for Historical Comparison”, *Computational Linguistics*, vol. 22, n° 4, p. 481-496, 1996.

Damerau F. J., “Mechanization of Cognate Recognition in Comparative Linguistics”, *Linguistics*, vol. 148, p. 5-29, 1975.

Dunning T., “Accurate Methods for the Statistics of Surprise and Coincidence”, *Computational Linguistics*, vol. 19, n° 1, p. 61-74, 1993.

Dyen I., Kruskal J. B., Black P., “An Indoeuropean Classification: A Lexicostatistical Experiment”, *Transactions of the American Philosophical Society*, 1992.

Eastlack C. L., “Iberochange: a program to simulate systematic sound change in Ibero-Romance”, *Computers and the Humanities*, vol. 11, p. 81-88, 1977.

Fellbaum C. (ed.), *WordNet: an electronic lexical database*, MIT, 1998.

Greenberg J. H., “Observations concerning Ringe’s *Calculating the factor of chance in language comparison*”, *Proc. of the American Philosophical Society*, vol. 137, p. 79-89, 1993.

Guy J. B. M., “An algorithm for identifying cognates in bilingual wordlists and its applicability to machine translation”, *Journal of Quantitative Linguistics*, vol. 1, n° 1, p. 35-42, 1994.

Hartman S. L., “A universal alphabet for experiments in comparative phonology”, *Computers and the Humanities*, vol. 15, p. 75-82, 1981.

Hewson J., “Comparative reconstruction on the computer”, *Proc. of the 1st International Conference on Historical Linguistics*, p. 191-197, 1974.

Hewson J., *A computer-generated dictionary of proto-Algonquian*, Canadian Museum of Civilization, 1993.

Johnson M., “Computer aids for comparative dictionaries”, *Linguistics*, vol. 23, n˚ 2, p. 285-302, 1985.

Kay M., The logic of cognate recognition in historical linguistics, Memorandum n˚ RM-4224-PR, The RAND Corporation, 1964.

Kessler B., “Computational dialectology in Irish Gaelic”, *EACL*, p. 60-67, 1995.

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

Knight K., Graehl J., “Machine Transliteration”, *Computational Linguistics*, vol. 24, n˚ 4, p. 599-612, 1998.

Koehn P., *Statistical Machine Translation*, Cambridge University Press, 2009. To appear.

Kondrak G., “A New Algorithm for the Alignment of Phonetic Sequences”, *NAACL*, p. 288-295, 2000.

Kondrak G., “Identifying Cognates by Phonetic and Semantic Similarity”, *NAACL*, p. 103-110, 2001.

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

Kondrak G., “Determining Recurrent Sound Correspondences by Inducing Translation Models”, *COLING*, p. 488-494, 2002b.

Kondrak G., “Identifying Complex Sound Correspondences in Bilingual Wordlists”, *CICLing*, p. 432-443, 2003a.

Kondrak G., “Phonetic Alignment and Similarity”, *Computers and the Humanities*, vol. 37, n˚ 3, p. 273-291, 2003b.

Kondrak G., “Combining Evidence in Cognate Identification”, *Canadian AI*, p. 44-59, 2004.

Kondrak G., Beck D., Dilts P., “Creating a Comparative Dictionary of Totonac-Tepehua”, *Proc. of the Workshop on Computing and Historical Phonology*, p. 134-141, 2007.

Kondrak G., Dorr B., “Automatic Identification of Confusable Drug Names”, *Artificial Intelligence in Medicine*, vol. 36, n˚ 1, p. 29-42, 2006.

Kondrak G., Marcu D., Knight K., “Cognates Can Improve Statistical Translation Models”, *HLT-NAACL*, p. 46-48, 2003.

Ladefoged P., *A Course in Phonetics*, Harcourt Brace Jovanovich, 1975.

Leacock C., Chodorow M., “Combining Local Context and WordNet Similarity for Word Sense Identification”, *in* C. Fellbaum (ed.), *WordNet: an electronic lexical database*, MIT, p. 265-283, 1998.

Lowe J. B., Mazaudon M., “The reconstruction engine: a computer implementation of the comparative method”, *Computational Linguistics*, vol. 20, p. 381-417, 1994.

Mann G. S., Yarowsky D., “Multipath Translation Lexicon Induction via Bridge Languages”, *NAACL*, p. 151-158, 2001.

Melamed I. D., “Automatic Discovery of Non-Compositional Compounds in Parallel Data”, *EMNLP*, p. 97-108, 1997.

Melamed I. D., “Bitext Maps and Alignment via Pattern Recognition”, *Computational Linguistics*, vol. 25, n˚ 1, p. 107-130, 1999.

Melamed I. D., “Models of Translational Equivalence among Words”, *Computational Linguistics*, vol. 26, n˚ 2, p. 221-249, 2000.

Myers E. W., “Seeing Conserved Signals”, *in* E. S. Lander, M. S. Waterman (eds), *Calculating the Secrets of Life*, National Academy Press, p. 56-89, 1995.

Nerbonne J., Heeringa W., “Measuring Dialect Distance Phonetically”, *SIGPHON*, 1997.

Oakes M. P., “Computer Estimation of Vocabulary in Protolanguage from Word Lists in Four Daughter Languages”, *Journal of Quantitative Linguistics*, vol. 7, n˚ 3, p. 233-243, 2000.

Oommen B. J., “String Alignment With Substitution, Insertion, Deletion, Squashing, and Expansion Operations”, *Information Sciences*, vol. 83, p. 89-107, 1995.

Oswalt R. L., “A probabilistic evaluation of North Eurasiatic Nostratic”, *in* J. C. Salmons, B. D. Joseph (eds), *Nostratic: sifting the evidence*, John Benjamins, p. 199-216, 1998.

Raman A., Newman J., Patrick J., “A Complexity Measure for Diachronic Chinese Phonology”, *SIGPHON*, 1997.

Ringe D., “On calculating the factor of chance in language comparison”, *Transactions of the American Philosophical Society*, 1992.

Ringe D., “Probabilistic Evidence for Indo-Uralic”, *in* J. C. Salmons, B. D. Joseph (eds), *Nostratic: sifting the evidence*, John Benjamins, p. 153-197, 1998.

Simard M., Foster G. F., Isabelle P., “Using Cognates to Align Sentences in Bilingual Corpora”, *TMI*, p. 67-81, 1992.

Smith R. N., “A computer simulation of phonological change”, *ITL: Tijdschrift voor Toegepaste Linguistiek*, vol. 1, n˚ 5, p. 82-91, 1969.

Smith T. F., Waterman M. S., “Identification of common molecular sequences”, *Journal of Molecular Biology*, vol. 147, p. 195-197, 1981.

Swadesh M., “Lexico-statistical dating of prehistoric ethnic contacts”, *Proc. of the American Philosophical Society*, vol. 96, p. 452-463, 1952.

Tiedemann J., “Automatic Construction of Weighted String Similarity Measures”, *EMNLP*, 1999.

Trask R. L., *Historical Linguistics*, Arnold, 1996.

Vossen P. (ed.), *EuroWordNet: a Multilingual Database with Lexical Semantic Networks*, Kluwer Academic, 1998.

Wagner R. A., Fischer M. J., “The String-to-String Correction Problem”, *Journal of the ACM*, vol. 21, n˚ 1, p. 168-173, 1974.

Watkins C. (ed.), *The American Heritage Dictionary of Indo-European Roots*, second edn, Houghton Mifflin, 2000.