Combining Evidence in Cognate Identification

Grzegorz Kondrak

Department of Computing Science,
University of Alberta,
Edmonton, AB, T6G 2E8, Canada
kondrak@cs.ualberta.ca
http://www.cs.ualberta.ca/~kondrak

Abstract. Cognates are words of the same origin that belong to distinct languages. The problem of automatic identification of cognates arises in language reconstruction and bitext-related tasks. The evidence of cognation may come from various information sources, such as phonetic similarity, semantic similarity, and recurrent sound correspondences. I discuss ways of defining the measures of the various types of similarity and propose a method of combining then into an integrated cognate identification program. The new method requires no manual parameter tuning and performs well when tested on the Indoeuropean and Algonquian lexical data.

1 Introduction

Cognates are words of the same origin that belong to distinct languages. For example, French *lait*, Spanish *leche*, and Italian *latte* constitute a set of cognates, since they all derive from Latin *lactem*. In general, the number of cognates between related languages decreases with time, and the ones that remain become less similar. Recurrent sound correspondences, which are produced by regular sound changes, are helpful in distinguishing cognate pairs from accidental resemblances. For example, the fact that /d/:/t/ is a recurrent correspondence between Latin and English (*ten/decem*, *tooth/dentem*. etc.) indicates that Latin *die* 'day' is not cognate with English *day*.

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 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 measure can be computed for any pair of words in isolation (levels 1, 2 and 3), but a longer list of related words is necessary for the determination of

1.	‘all’	alə	ɟiθə
2.	‘and’	unt	e
3.	‘animal’	tīr	kafšə
4.	‘ashes’	ašə	hi
5.	‘at’	an	nə
6.	‘back’	rükən	špinə
7.	‘bad’	šlext	kec
8.	‘bark’	rində	škəlbozə
9.	‘because’	vayl	sepse
10.	‘belly’	bawx	bark

Table 1. An excerpt from the German/Albanian word-pair list [10].

the recurrent sound correspondences (levels 2 and 3), while a semantic measure is only applicable when words are accompanied by glosses (level 3).

The ultimate goal of the research described in this paper is the fascinating possibility of performing an automatic reconstruction of proto-languages from the information contained in the descendant languages. Given dictionaries of related languages, a hypothetical language reconstruction program would be able to determine recurrent sound correspondences, identify cognate sets, and reconstruct their corresponding proto-forms.

The identification of cognates is not only the key issue in language reconstruction, but is also important in a number of bitext-related tasks, such as sentence alignment [3, 19, 21, 24], inducing translation lexicons [11, 18], and improving statistical machine translation models [1]. Most of the applications take advantage of the fact that nearly all co-occurring cognates in bitexts are mutual translations. In the context of bitexts, the term *cognate* usually denotes words in different languages that are similar in form and meaning, without making a distinction between borrowed and genetically related words.

Current approaches to cognate identification employ either phonetic/orthographic similarity measures [2, 19, 21, 23] or recurrent sound/letter correspondences [6, 18, 26]. However, there have been very few attempts to combine dif-

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

Table 2. Excerpts from the Cree (left) and the Ojibwa (right) vocabulary lists [9].

ferent ways of cognate identification. Yarowsky and Wincentowski [28] bootstrap the values of edit cost matrix with rough phonetic approximations, and then iteratively re-estimate the matrix in order to derive empirically observed character-to-character probabilities. Kondrak [13] linearly combines a phonetic score with a semantic score of gloss similarity.

In this paper, I present a method of integrating distinct types of evidence for the purpose of cognate identification. In particular, the combined phonetic and correspondence-based similarity measures are applied to lists of word pairs, and the semantic similarity of glosses is added on when dealing with vocabulary lists. The new method combines various similarity scores in a principled way. In terms of accuracy, when tested on independently compiled word and vocabulary lists, it matches or surpasses the results obtained using the method with manually set parameters [13]. Finally, the method makes it possible to utilize complex, multi-phoneme correspondences for cognate identification.

The paper is organized as follows. The next three sections provide background on the measures of phonetic, correspondence-based, and semantic similarity, respectively, in the context of cognate identification. After introducing the method of combining various measures, I describe and discuss experimental results on authentic language data. I conclude with a comparison of the method presented here with another method of identifying cognates.

2 Phonetic Similarity

Surface-form similarity of words can be estimated using orthographic and/or phonetic measures. Simple measures of orthographic similarity include edit distance [19], Dice's bigram similarity coefficient [2], and the Longest Common Subsequence Ratio (LCSR) [21]. Phonetic measures are applicable if words are given in a phonetic or phonemic transcription. ALINE [12] is a phonetic word aligner based on multivalued phonetic features with salience weights. Thanks to its ability to assess the similarity of individual segments, ALINE performs better on cognate identification than the orthographic measures that employ a binary identity function on the level of character comparison [13].

ALINE returns a normalized score in the $[0, 1]$ range. The score by itself can be used to rank candidate pairs with respect to their phonetic similarity. However, in order to combine the phonetic score with the semantic and/or correspondence-based scores, it is helpful to convert the score assigned to a pair of words into the probability that they are related. For modeling the distribution of scores, I adopt the Beta distribution. The Beta distribution is defined over the domain $[0, 1]$, and has two free parameters $\hat{A}$ and $\hat{B}$. The relationship between the two parameters and the mean and variance of the distribution is the following:

$$\mu = \frac{\hat{A}}{\hat{A} + \hat{B}} \qquad\qquad \sigma^2 = \frac{\hat{A}\hat{B}}{(\hat{A} + \hat{B})^2(\hat{A} + \hat{B} + 1)}$$

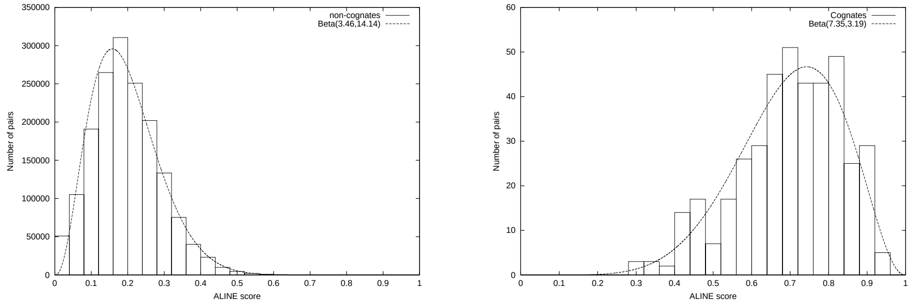

Fig. 1. Distribution of the phonetic scores for the unrelated (left) and the cognate (right) word pairs, and the corresponding Beta distributions.

Figure 1 shows the distribution of phonetic scores between word pairs in the development set (Cree–Ojibwa) within the 0.04 intervals. The left and right plot depict the phonetic score distribution for the unrelated and for the cognate word pairs, respectively. The parameters of the corresponding Beta distributions were calculated from the mean and variance of the scores using the relationship expressed in formulas (1) and (2). For unrelated words, the Beta distribution fits the distribution of phonetic scores remarkably well. For cognate words, the fit is also quite good although 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.

3 Correspondence-Based Similarity

3.1 Determination of Simple Correspondences

For the determination of recurrent sound correspondences (often referred to simply as correspondences) I employ the method of inducing a *translation model* between phonemes in two wordlists [14]. The idea is to relate recurrent sound correspondences in wordlists to translational equivalences in bitexts. The translation model is induced by combining the maximum similarity alignment with the competitive linking algorithm of Melamed [22]. Melamed's approach is based on the *one-to-one* assumption, which implies that every word in the bitext is aligned with at most one word on the other side of the bitext. In the context of the bilingual wordlists, the correspondences determined under the *one-to-one* assumption are restricted to link single phonemes to single phonemes. Nevertheless, the method is powerful enough to determine valid correspondences in wordlists in which the fraction of cognate pairs is well below 50% [14].

The correspondence-based similarity score between two words is computed in the following way. Each valid correspondence is counted as a link and contributes a constant positive score (no crossing links are allowed). Each unlinked segment, with the exception of the segments beyond the rightmost link, is assigned a smaller negative score. The alignment with the highest score is found using dynamic programming [27]. If more than one best alignment exists, links are assigned the weight averaged over the entire set of best alignments. Finally, the score is normalized by dividing it by the average of the lengths of the two words.

3.2 Determination of Complex Correspondences

Kondrak [15] proposed an extension of the one-to-one method that is capable of discovering complex, ‘many-to-many” correspondences. The method is an adaptation of Melamed’s algorithm for discovering non-compositional compounds in bitexts [20]. A non-compositional compound (NCC) is a word sequence, such as “high school”, whose meaning cannot be synthesized from the meaning of its components. Experimental results indicate that the method can achieve up to 90% recall and precision in determination of correspondences on vocabulary lists [15].

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 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.

Figure 2 shows the distribution of correspondence-based scores between word pairs in the development set (Cree–Ojibwa). For unrelated words, the fit with the Beta distribution is not as good as in the case of phonetic scores, but still acceptable. For cognate words, the Beta distribution fails to account for a number of word pairs that are perfectly covered by correspondences (score = 1.0). However, the problem is likely to be less acute for language pairs that are not as closely related as Cree and Ojibwa.

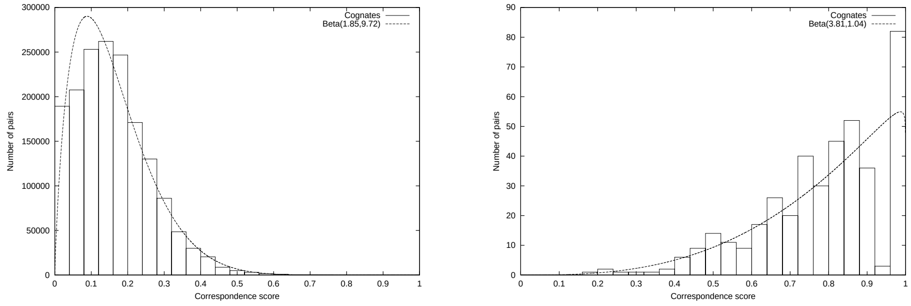

Fig. 2. Distribution of the correspondence-based scores for the unrelated (left) and the cognate (right) word pairs, and the corresponding Beta distributions.

3.3 Determination of Correspondences in Vocabulary Lists

Kondrak [14] showed that correspondences determined with the one-to-one approach can be successfully used for cognate identification in pairs of word lists, where the words are already matched by meanings. However, the task is more challenging when cognates have to be identified in unstructured vocabulary lists, In vocabulary lists, as opposed to word lists, words across languages are not neatly matched by meanings; rather, semantic similarity has to be inferred from glosses. Whereas in word lists of related languages the percentage of cognates can be expected to exceed 10%, the probability that randomly selected words from two vocabulary lists are cognate is usually less than 0.1%. Attempting to determine correspondences with such a small signal-to-noise ratio is bound to fail. It is necessary first to identify a smaller set of likely cognate pairs, on which a translation model can be successfully induced.

One possible way to determine the set of likely cognate pairs is to select n candidate pairs starting from the top of the ordered list produced by a combined semantic and phonetic approach. However, the selected pairs are likely to include many pairs that exhibit high phonetic similarity. When a translation model is induced on such set, the strongest correspondences can be expected to consist mostly of pairs of identical phonemes.

A better approach, which is not biased by the phonetic similarities between phonemes, is to select candidate pairs solely on the basis of semantic similarity. The idea is to extract all vocabulary entries characterized by the highest level of semantic similarity, that is, gloss identity. Even though such a set is still likely to contain mostly unrelated word pairs, the fraction of cognates may be sufficiently large to determine the strongest correspondences. The determined correspondences can then be used to identify cognates among all possible word pairs.

4 Semantic Similarity Features

Kondrak [13] developed a scheme for computing semantic similarity of glosses on the basis of keyword selection and WordNet [5] lexical relations. The scheme combines four lexical relations and two focus levels, which together yield eight semantic similarity levels (Table 3). Keywords are salient words extracted from glosses by a heuristic method based on part-of-speech tags. If there exists a lexical relationship in WordNet linking the two glosses or any of their keywords, the semantic similarity score is determined according to the scheme shown in Table 3. The levels of similarity are considered in descending score order, with keyword identity taking precedence over gloss hypernymy. The scores are not cumulative. The numerical values in Table 3 were set manually on the basis of experiments with the development set (Cree–Ojibwa).

In this paper, I propose to consider the eight semantic similarity levels as binary semantic *features*. Although the features are definitely not independent, it may be advantageous to consider their combinations rather than just simply

Lexical relation	Focus level	
	Gloss	Keyword
Identity	1.00	0.50
Synonymy	0.70	0.35
Hypernymy	0.50	0.25
Meronymy	0.10	0.05

Table 3. Semantic similarity features and their numerical scores [13].

the most prominent one. For example, gloss hypernymy accompanied by keyword synonymy might constitute stronger evidence for a semantic relationship than gloss hypernymy alone.

In the context of detecting semantic similarity of glosses, a transitive *subsumption* relation can be defined for the semantic features. In the following assertions, the expression "feature A subsumes feature B" should be understood as "feature B is redundant if feature A is present".

1. Gloss identity subsumes other relations involving glosses (e.g. gloss identity subsumes gloss meronymy).
2. Keyword identity subsumes other relations involving keywords.
3. Features involving a lexical relation between glosses subsume features involving the same lexical relation between keywords (e.g. gloss hypernymy subsumes keyword hypernymy).
4. Synonymy subsumes hypernymy and meronymy, and hypernymy subsumes meronymy.

The resulting partial ordering of features is shown in Figure 3. Assertion 4 is probably the most debatable.

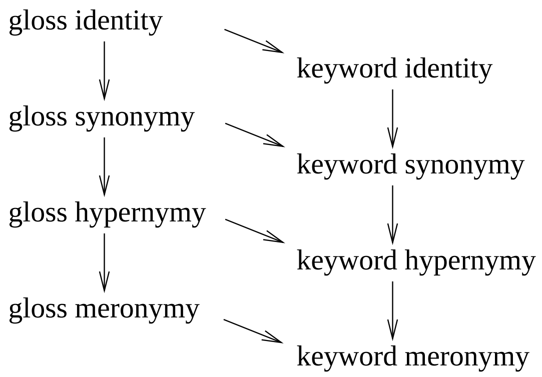

Fig. 3. Partial ordering of semantic features.

sem	=	related to semantic similarity
ph	=	related to phonetic similarity
rc	=	related to correspondence-based similarity
$+$	=	related to cognate pairs
$-$	=	related to unrelated pairs
$cogn$	=	the given pair of words are cognate
$\neg cogn$	=	the given pair of words are unrelated
v	=	feature vector for the given word pair
v_j	=	values of the binary semantic features
s	=	numerical scores for the given word pair (in the $[0,1]$ range)
σ	=	partial similarity scores
α	=	interpolation parameters (weights)
d	=	probability density function of the corresponding Beta distribution
C_i	=	normalizing constants independent of the given word pair

Table 4. Symbols used in Section 5.

I investigated the following variants of semantic feature ordering:

LN The linear order that corresponds to the semantic scale originally proposed in [13].
SP The partial order implied by assertions 1–4, which is shown in Figure 3.
WP The weaker version of the partial order, implied by assertions 1–3.
NO The unordered set of the eight semantic features.
MK The feature set corresponding to Method K in [13], which contains only the WordNet-independent features: gloss identity and keyword identity (the former subsumes the latter).
MG The feature set corresponding to Method G in [13], which contains only gloss identity.
NS Empty set, i.e. no semantic features are used (the baseline method).

I discuss the effect of the feature ordering variants on the overall accuracy in Section 6.

5 Combining Various Types of Evidence

In [13], the overall similarity score was computed using a linear combination of the semantic and the phonetic scores. The interpolation parameter was determined on a development set. In this paper, I adopt the Naive Bayes approach to combining various sources of information. The vector v consists of the eight semantic features, the phonetic similarity score, and the correspondence-based similarity score. The overall word-pair similarity score for a pair of words is computed by the following formula (see Table 4 for the explanation of symbols):

$$score = \frac{p(cogn \mid v)}{p(\neg cogn \mid v)} = \frac{p(cogn) \cdot p(v \mid cogn)}{p(\neg cogn) \cdot p(v \mid \neg cogn))} = C_1 \cdot \frac{p(v \mid cogn)}{p(v \mid \neg cogn))}$$

$$= C_2 \cdot \left(\frac{(\prod_j p(v_j \mid cogn))}{(\prod_j p(v_j \mid \neg cogn))} \right)^{\alpha_{sem}} \cdot \left(\frac{d_{ph+}(s_{ph})}{d_{ph-}(s_{ph})} \right)^{\alpha_{ph}} \cdot \left(\frac{d_{rc+}(s_{rc})}{d_{rc-}(s_{rc})} \right)^{\alpha_{rc}}$$

It is more convenient to do the computations using logarithms:

$$\log score = \alpha_{sem} \cdot \sigma_{sem} + \alpha_{ph} \cdot \sigma_{ph} + \alpha_{rc} \cdot \sigma_{rc} + C_3$$

where

$$\sigma_{sem} = \sum\nolimits_j \log \frac{p(v_j \mid cogn)}{p(v_j \mid \neg cogn)}, \qquad \sigma_{ph} = \log \frac{d_{ph+}(s_{ph})}{d_{ph-}(s_{ph}))}, \qquad \sigma_{rc} = \log \frac{d_{rc+}(s_{rc})}{d_{rc-}(s_{rc})}.$$

Since the goal is a relative ranking of candidate word-pairs, the exact value of the normalizing constant C_3 is irrelevant.

The difference between the current method of combining partial scores and the method presented in [13] lies in the way the original scores are transformed into probabilities using the Naive Bayes assumption and the Beta distribution. A number of parameters must still be established on a separate training set: the conditional probability distributions of the semantic features, the parameters for the beta distributions, and the interpolation parameters. However, the values of the parameters (except the interpolation parameters) are set automatically rather than manually.

6 Results

The cognate identification methods were tested on two different data sets: a set of structured word lists of 200 basic meanings, and a set of unstructured vocabulary lists containing thousands of entries with glosses.

The accuracy of the methods was evaluated by computing the 11-point interpolated average precision for the vocabulary lists, and the n-point average precision for the word lists (n is the total number of cognate pairs in a list). The output of the system is a list of suspected cognate pairs sorted by their similarity scores. Typically, true cognates are very frequent near the top of the list, and become less frequent towards the bottom. The threshold value that determines the cut-off depends 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 threshold, precision is computed at a number of different recall levels, and then averaged to yield a single number. In the case of the 11-point average precision, the recall levels are set at 0%, 10%, 20%, ..., 100%. In the case of the n-point average precision, precision is calculated at each point in the list where a true cognate pair is found. In the experiments reported below, I uniformly assumed the precision value at 0% recall to be 1, and the precision value at 100% recall to be 0.

6.1 Results on the Indoeuropean Word Lists

The experiments in this section were performed using a list of 200 basic meanings that are considered universal and relatively resistant to lexical replacement [25].

Languages		Phonetic	Simple	Complex	Phonetic + Simple	Phonetic + Complex
English	German	.916	.949	.924	.946	.930
French	Latin	.863	.869	.874	.881	.882
English	Latin	.725	.857	.740	.828	.796
German	Latin	.706	.856	.795	.839	.830
English	French	.615	.557	.556	.692	.678
French	German	.504	.525	.526	.575	.572
Albanian	Latin	.618	.613	.621	.696	.659
Albanian	French	.612	.443	.460	.600	.603
Albanian	German	.323	.307	.307	.395	.398
Albanian	English	.277	.202	.243	.340	.330
Average		**.616**	**.618**	**.605**	**.679**	**.668**

Table 5. The average cognate identification precision on the Indoeuropean 200-word lists for various methods.

The development set included six 200-word lists (Italian, Polish, Romanian, Russian, Serbocroatian and Spanish) adapted from the Comparative Indoeuropean Data Corpus [4]. The test set consisted of five lists (Albanian, English, French, German, and Latin) compiled by Kessler [10]. In this experiment, only words belonging to the same semantic slot were considered as possible cognates.

Table 5 compares the average cognate identification precision on the test set obtained using the following methods:

Phonetic The phonetic approach, in which cognate pairs are ordered according to their phonetic similarity score computed by ALINE. The settings of ALINE's parameters are the same as in [12].

Simple The correspondence-based approach, as described in [14] (method D), in which only simple, one-to-one correspondences are identified.

Complex The correspondence-based approach that identifies complex, many-to-many correspondences [15].

Phonetic + Simple The combination of the phonetic and the correspondence-based approaches, without utilizing complex correspondences.

Phonetic + Complex The combination of the phonetic and the correspondence-based approaches that utilizes complex correspondences.

In the final two variants, the phonetic and the correspondence-based approaches are combined using the method described in Section 5, with the parameters derived from the Italian/Polish word list (the interpolation parameters α_{ph} and α_{rc} were held equal to 1). This particular language pair was chosen because it produced the best overall results on the development set. However, the relative differences in average precision with different training sets did not exceed 1%.

The results in Table 5 show that both the phonetic method and the correspondence-based method obtain similar average cognate identification precision. The combination of the two methods achieves a significantly higher precision.

Languages		NS	MG	MK	LN	SP	WP	NO
Fox	Menomini	.488	.607	.640	.651	.652	.652	.491
Fox	Cree	.508	.682	.678	.698	.694	.694	.549
Fox	Ojibwa	.655	.674	.685	.691	.695	.695	.572
Menomini	Cree	.438	.591	.612	.618	.613	.608	.523
Menomini	Ojibwa	.478	.611	.632	.641	.639	.635	.516
Average on test set		**.513**	**.633**	**.649**	**.660**	**.658**	**.657**	**.530**
Cree	Ojibwa	.717	.783	.785	.787	.784	.784	.722

Table 6. The average precision on the Algonquian vocabulary lists obtained by combining the semantic similarity features, the phonetic similarity score, and the complex-correspondence-based similarity score.

Surprisingly, the incorporation of complex correspondences 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 Indoeuropean 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.

Additional experiments showed 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 5. On the test set, the straightforward method achieves the average precision of .683 with simple correspondences, and .680 with complex correspondences.

6.2 Results on the Algonquian Vocabulary Lists

The cognate identification method was also tested on noun portions of four Algonquian vocabulary lists [9]. The lists representing Fox, Menomini, Cree, and Ojibwa contain over 4000 noun entries in total. The results were evaluated against an electronic version of the Algonquian etymological dictionary [8]. The dictionary contains 4,068 cognate sets, including 853 marked as nouns. The Cree–Ojibwa language pair was used as the development set, while the remaining five pairs served as the test set. The proportion of cognates in the set of word-pairs that have at least one gloss in common was 33.1% in the development set and ranged from 17.5% to 26.3% in the test set.

Table 6 shows the average precision obtained on the Algonquian data by combining the phonetic, semantic and (complex) correspondence-based similarity using the method presented in Section 5. The columns correspond to variants of semantic feature ordering defined in Section 4. The numbers shown in bold type in the left-most four columns can be directly compared to the corresponding results obtained using the method described in [13]. The latter method, which uses the linear combination of the phonetic and the semantic similarity scores (set according to Table 3), achieved the 11-point average precision of .430, .596, .617, and .628, for variants NS, MG, MK, and LN, respectively. Therefore,

Methods	Correspondences		
	None	Simple	Complex
—	—	.448	.473
Phonetic	.430	.472	.513
Semantic	.227	.633	.625
Phonetic + Semantic	.631	.652	.660

Table 7. The average precision on the Algonquian vocabulary lists (test set only) obtained by combining various methods.

the improvement ranges from 5% (LN) to nearly 20% (NS). When all semantic features are utilized (columns LN, SP, WP, and NO), there is hardly any difference in average precision between alternative orderings of semantic features (LN, SP, WP). However, applying the features without any ordering (NO) is almost equivalent to using no semantics at all (NS).

Table 7 provides more details on the contribution of various types of evidence to the overall average precision. For example, the merger of the phonetic and the semantic similarity with no recourse to correspondences achieves the average precision of .631. (not significantly better than the average precision of .628 obtained using the method described in [13]). Replacing the phonetic similarity with the (simple) correspondence-based similarity has little influence on the average precision: .448 vs .430 without semantics, and .633 vs. .631 with semantics. The advantage provided by complex correspondences all but disappears when all types of evidence are combined (.660 vs. .652). Relying on gloss similarity alone is inadequate (.227) because no continuous score is available to order candidate pairs within the semantic similarity classes.

The tests were performed with the following parameter settings: semantic feature ordering – linear (LN); parameters for computing phonetic similarity – as in [13]; parameters for computing the correspondence-based score – as in [14] (complex correspondences limited to consonant clusters); number of iterations of the NCC algorithm — 12, as in [15]. When two types of evidence were combined, the interpolation parameters were held equal to 1. With all three types of evidence, the interpolation parameters were $\alpha_{sem} = 2$, $\alpha_{ph} = 1$, and $\alpha_{rc} = 1$.

The choice of values for the interpolation parameters requires further explanation. The weights used for the final testing were selected because they are relatively simple and result in near-maximum average precision on the training data. They also have a theoretical justification. Both the phonetic and the correspondence-based similarity measures are calculated on the basis of the phonetic transcription of words. Moreover, recurrent correspondences are composed mostly of similar or identical phonemes. In contrast, the semantic similarity measure is based exclusively on glosses. The experiments performed on the training set suggested that the best results are obtained by assigning approximately equal weight to the gloss-based evidence and to the lexeme-based measures. The results in Tables 7 and 6 reflect this observation. If the weights are equal for all three types of evidence, the average precision drops to .616 with the simple correspondences, and to .639 with the complex correspondences.

7 Computing Similarity vs. Generating Proto Projections

It is interesting to compare the method described here to the method that was originally used to compile the etymological dictionary [8], which served as our gold standard, from the vocabulary lists [9], which also constituted our test data in Section 6.2. The method [7] is based on generating proto-projections (candidate proto-forms) of the lexemes occurring in the vocabulary lists of the daughter languages. For example, assuming that the consonant cluster *šš* in Ojibwa is a reflex of either **hš* or **qš* in Proto-Algonquian, the proto-projections of Ojibwa *mišši* 'piece of firewood' would include **mihši* and **miqši.* The cognate identification process succeeds if the intersection of the sets of proto-projections generated from distinct daughter languages is not empty. The set intersection operation was implemented by alphabetically sorting all proto-projections. The potential cognate sets were subsequently analyzed by a linguist in order to determine whether they were in fact reflexes of the same proto-form and, if that was the case, to reconstruct the proto-form.

Hewson's method has a number of disadvantages. It is based exclusively on recurrent sound correspondences, with no recourse to potentially valuable phonetic and semantic information. It requires the user to provide a complete table of correspondences between daughter languages and the reconstructed proto-language. Since such a table of correspondences is established on the basis of multiple sets of confirmed cognates, the method is applicable only to language families that have already been throughly analyzed. In addition, the number of proto-projections increases combinatorially with the number of ambiguous reflexes that occur in a word. Anything less than a perfect match of correspondences may result in a cognate pair being overlooked.

Table 8 contains some interesting examples of Algonquian cognate pairs that are not found in Hewson's dictionary, but are recognized by the implementation of the method proposed in this paper. Their semantic, phonetic, and correspondence-based similarity considered in isolation may not be sufficient for their identification, but combining all three types of evidence results in a high overall similarity score. In particular, such pairs are bound to be missed by any approach that requires the identity of glosses as the necessary condition for consideration.

8 Conclusion

I have proposed a method of combining various types of evidence for the task of automatic cognate identification. In many cases, the new method achieves higher accuracy than the method based on the linear combination of scores. Moreover, the new method does not require manual parameter tuning, but instead can be trained on data from other language pairs.

#	Lang.	Lexeme	Gloss	WordNet relation
1	Cree	*mōsāpēw*	‘unmarried man’	*gloss synonymy*
	Men.	*mōsāpēwew*	‘bachelor, single man’	
2	Fox	*kešēmanetōwa*	‘great spirit’	*none*
	Men.	*kesēmanetōw*	‘god’	
3	Cree	*wīhkēs*	‘sweet-flag’	*keyword hypernymy*
	Ojib.	*wīkkēn’*	‘iris’	
4	Men.	*enōhekan*	‘pointer’	*none*
	Ojib.	*inō’ikan*	‘that which is pointed at’	
5	Fox	*mīkātiweni*	‘fight’	*gloss synonymy*
	Men.	*mīkātwan*	‘war, fighting’	
6	Fox	*atāmina*	‘maize-plant’	*keyword synonymy*
	Ojib.	*mantāmin*	‘grain of corn’	
7	Fox	*ātesōhkākana*	‘sacred story’	*keyword identity*
	Ojib.	*ātissōkkān*	‘story or legend’	

Table 8. Examples of cognate pairs not included in Hewson’s dictionary.

The method proposed here is applicable both to structured (word lists) and unstructured (vocabulary lists) data. Apart from assisting the comparative linguists in proto-language reconstruction, it can be used to dramatically speed up the process of producing etymological dictionaries, even when little is known about the languages in question. The results of the experiments show that it is possible to discover a large number of cognates with good precision. To take the Fox–Menomini pair as an example, 70% recall at 50% precision signifies that the top 170 candidates contain 85 out of 121 existing cognate pairs. Moreover, many of the apparent false positives are in fact cognates or lexemes that are related in some way.

This paper belongs to a line of research that has already resulted in applications in such diverse areas as statistical machine translation [17] and the identification of confusable drug names [16]. In the long run, such applications may prove 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.

Acknowledgments

Thanks to Colin Cherry, Russ Greiner, Dekang Lin, Dan Lizotte, and Dale Schuurmans for comments and advice. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

1. Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. Statistical machine translation. Technical report, Johns Hopkins University, 1999.
2. Chris Brew and David McKelvie. Word-pair extraction for lexicography. In K. Oflazer and H. Somers, editors, *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55, Ankara, Bilkent University, 1996.
3. Kenneth W. Church. Char_align: A program for aligning parallel texts at the character level. In *Proceedings of ACL-93: 31st Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Columbus, Ohio, 1993.
4. Isidore Dyen, Joseph B. Kruskal, and Paul Black. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5), 1992.
5. Christiane Fellbaum, editor. *WordNet: an electronic lexical database.* The MIT Press, Cambridge, MA, 1998.
6. Jacques B. M. Guy. An algorithm for identifying cognates in bilingual wordlists and its applicability to machine translation. *Journal of Quantitative Linguistics*, 1(1):35–42, 1994. MS-DOS executable available at http://garbo.uwasa.fi.
7. John Hewson. Comparative reconstruction on the computer. In *Proceedings of the 1st International Conference on Historical Linguistics*, pages 191–197, 1974.
8. John Hewson. *A computer-generated dictionary of proto-Algonquian.* Hull, Quebec: Canadian Museum of Civilization, 1993.
9. John Hewson. Vocabularies of Fox, Cree, Menomini, and Ojibwa, 1999. Computer file.
10. Brett Kessler. *The Significance of Word Lists.* Stanford: CSLI Publications, 2001. Word lists available at http://spell.psychology.wayne.edu/~bkessler.
11. Philipp Koehn and Kevin Knight. Knowledge sources for word-level translation models. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 27–35, 2001.
12. Grzegorz Kondrak. A new algorithm for the alignment of phonetic sequences. In *Proceedings of NAACL 2000: 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 288–295, 2000.
13. Grzegorz Kondrak. Identifying cognates by phonetic and semantic similarity. In *Proceedings of NAACL 2001: 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 103–110, 2001.
14. Grzegorz Kondrak. Determining recurrent sound correspondences by inducing translation models. In *Proceedings of COLING 2002: 19th International Conference on Computational Linguistics*, pages 488–494, 2002.
15. Grzegorz Kondrak. Identifying complex sound correspondences in bilingual wordlists. In *Proceedings of CICLing 2003: 4th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 432–443, 2003.
16. Grzegorz Kondrak and Bonnie Dorr. Identification of confusable drug names: A new approach and evaluation methodology. 2004. In preparation.
17. Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. Cognates can improve statistical translation models. In *Proceedings of HLT-NAACL 2003: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–48, 2003. Companion volume.

18. Gideon S. Mann and David Yarowsky. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL 2001: 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 151–158, 2001.
19. Tony McEnery and Michael Oakes. Sentence and word alignment in the CRATER Project. In J. Thomas and M. Short, editors, *Using Corpora for Language Research*, pages 211–231. Longman, 1996.
20. I. Dan Melamed. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 97–108, 1997.
21. I. Dan Melamed. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130, 1999.
22. I. Dan Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.
23. Michael P. Oakes. Computer estimation of vocabulary in protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, 7(3):233–243, 2000.
24. Michel Simard, George F. Foster, and Pierre Isabelle. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, Montreal, Canada, 1992.
25. Morris Swadesh. Lexico-statistical dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96:452–463, 1952.
26. Jörg Tiedemann. Automatic construction of weighted string similarity measures. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, Maryland, 1999.
27. Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173, 1974.
28. David Yarowsky and Richard Wincentowski. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL-2000*, pages 207–216, 2000.