Identifying Cognates by Phonetic and Semantic Similarity

Grzegorz Kondrak
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada M5S 3G4
kondrak@cs.toronto.edu

Abstract

I present a method of identifying cognates in the vocabularies of related languages. I show that a measure of phonetic similarity based on multivalued features performs better than “orthographic” measures, such as the Longest Common Subsequence Ratio (LCSR) or Dice’s coefficient. I introduce a procedure for estimating semantic similarity of glosses that employs keyword selection and WordNet. Tests performed on vocabularies of four Algonquian languages indicate that the method is capable of discovering on average nearly 75% percent of cognates at 50% precision.

1 Introduction

In the narrow sense used in historical linguistics, cognates are words in related languages that have developed from the same ancestor word. An example of a cognate pair is French *lait* and Spanish *leche*, both of which come from Latin *lacte*. In other contexts, including this paper, the term is often used more loosely, denoting words in different languages that are similar in form and meaning, without making a 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 historical linguistics, the identification of cognates is a component of two principal tasks of the field: establishing the relatedness of languages and reconstructing the histories of language families. In corpus linguistics, cognates have been used for bitext alignment (Simard et al., 1992; Church, 1993; McEnery and Oakes, 1996; Melamed, 1999), and for extracting lexicographically interesting word-pairs from multilingual corpora (Brew and McKelvie, 1996).

The task addressed in this paper can be formulated in two ways. On the word level, given two words (*lexemes*) from different languages, the goal is to compute a value that reflects the likelihood of the pair being cognate. I assume that each lexeme is given in a phonetic notation, and that it is accompanied by one or more glosses that specify its meaning in a metalanguage for which a lexical resource is available (for example, English). On the language level, given two vocabulary lists representing two languages, the goal is to single out all pairs that appear to be cognate. Tables 1 and 2 show sample entries from two typical vocabulary lists. Such vocabulary lists are sometimes the only data available for lesser-studied languages.

In general, deciding whether two words are genetically related requires expert knowledge of the history of the languages in question. With time, words in all languages change their form and meaning. After several millennia, cognates often acquire very different phonetic shapes. For example, English *hundred*, French *cent*, and Polish *sto* are all descendants of Proto-Indo-European **kmtom*. 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 be a matter of chance resemblance, as in English *day* and Latin *die* ‘day’.

In the traditional approach to cognate identification, words with similar meanings are placed side by side. Those pairs that exhibit some phonological similarity are analyzed in order to find systematic correspondences of sounds. The correspondences in turn can be used to distinguish between genuine cognates and borrowings or chance resemblances.

My approach to the identification of cognates is based on the assumption that, in spite of the inevitable diachronic changes, cognates on average display higher semantic and phonetic similarity than

āniskōhōčikan	string of beads tied end to end
asikan	sock, stocking
kamāmakos	butterfly
kostāčīwin	terror, fear
misiyēw	large partridge, hen, fowl
namēhpin	wild ginger
napakihtak	board
tēhtēw	green toad
wayakēskw	bark

Table 1: An excerpt from a Cree vocabulary list (Hewson, 1999).

āšikan	dock, bridge
anaka’ēkkw	bark
kipaskosikan	medicine to induce clotting
kottāčīwin	fear, alarm
mēmīkwan’	butterfly
misissē	turkey
namēpin	sucker
napakissakw	plank
tēntē	very big toad

Table 2: An excerpt from an Ojibwa vocabulary list (Hewson, 1999).

words that are unrelated.[1] In this paper, I present COGIT, a cognate-identification system that combines ALINE (Kondrak, 2000), a feature-based algorithm for measuring phonetic similarity, with a novel procedure for estimating semantic similarity that employs keyword selection and WordNet. When tested on data from four native American languages, COGIT was able to discover, on average, nearly 75% percent of cognates at 50% precision, without resorting to a table of systematic sound correspondences. The results show that a large percentage of cognates can be detected automatically.

2 Related work

To my knowledge, no previously proposed algorithmic method is able to identify cognates directly in vocabulary lists. Guy's (1994) program COGNATE identifies probable letter correspondences between words and estimates how likely it is that the words are related. The algorithm has no semantic component, as the words are assumed to have already been matched by their meanings. Such an approach by definition cannot detect cognates that have undergone a semantic shift. Hewson (1974; 1993) employed a simple strategy of generating proto-projections to produce a dictionary of over 4000 Proto-Algonquian etyma from vocabularies of several contemporary Algonquian languages. The proto-projections, generated using long-established systematic sound correspondences, were then examined individually in order to select true cognates. The “Reconstruction Engine” of Lowe and Mazaudon (1994) uses a similar strategy of generating proto-projections to establish cognate sets. Both Hewson's and Lowe and Mazaudon's approaches require a complete table of systematic sound correspondences to be provided beforehand. Such tables can be constructed for well-studied language families on the basis of previously identified cognate sets, but are not available for many African and native American languages, especially in the cases where the relationship between languages has not been adequately proven. In contrast, the method presented in this paper operates directly on the vocabulary lists.

3 Phonetic similarity

The approaches to measuring word similarity can be divided into two groups. The “orthographic” approaches disregard the fact that alphabetic symbols express actual sounds, employing a binary identity function on the level of character comparison. A one-to-one encoding of symbols has no effect on the results. The “phonetic” approaches, on the other hand, attempt to take advantage of the phonetic characteristics of individual sounds in order to estimate their similarity. This presupposes a transcription of the words into a phonetic or phonemic representation.

The “orthographic” approaches are commonly used in corpus linguistics. Simard et al. (1992) consider two words to be cognates if their first four characters are identical (the “truncation” method). Brew and McKelvie (1996) use a number of methods based on calculating the number of shared character bigrams. For example, Dice's coefficient is defined as

$$DICE(x,y) = \frac{2|bigrams(x) \cap bigrams(y)|}{|bigrams(x)| + |bigrams(y)|}$$

where *bigrams(x)* is a multi-set of character bigrams in *x*. Church (1993) uses 4-grams at the level

[1] The assumption was verified during the evaluation of my system (Section 6). However, in the case of very remotely related languages, the difference may no longer be statistically significant (Ringe, 1998).

of character sequences. Melamed (1999) uses the Longest Common Subsequence Ratio (LCSR) defined as

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

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

ALINE (Kondrak, 2000), is an example of the “phonetic” approach. ALINE was originally developed for aligning phonetic sequences, but since it chooses the optimal alignment on the basis of a similarity score, it can also be used for computing similarity. Each phoneme is represented as a vector of phonetically-based feature values. The number of distinct values for each feature is not constrained.[2] The features have salience coefficients that express their relative importance. ALINE uses dynamic programming to compute similarity scores. 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.

Intuitively, a complex algorithm such as ALINE should be more accurate than simple, “orthographic” coefficients. By applying various methods to a specific task, such as cognate identification, their relative performance can be objectively evaluated.

4 Semantic similarity

The meanings of the lexemes are represented by their glosses. Therefore, the simplest method to detect semantic similarity is to check if the lexemes have at least one gloss in common. For example, the cognates *kottāčīwin* ‘terror, fear’ and *kostāčīwin* ‘fear, alarm’ in Tables 1 and 2 are correctly associated by this method. However, in many cases, the similarity of semantically related glosses is not recognized. The most common reasons are listed below.

1. Spelling errors or variants: ‘vermilion’ and ‘vermillion’, ‘sweet grass’ and ‘sweetgrass’, ‘plow’ and ‘plough’;
2. Morphological differences: ‘ash’ and ‘ashes’;
3. Determiners: ‘a mark’ and ‘mark’, ‘my finger’ and ‘finger’, ‘fish’ and ‘kind of fish’;
4. Adjectival modifiers: ‘small stone’ and ‘stone’;
5. Nominal modifiers: ‘goose’ and ‘snow goose’;
6. Complements and adjuncts: ‘stone’ and ‘stone of peach’, ‘island’ and ‘island in a river’;
7. Synonymy: ‘grave’ and ‘tomb’;
8. Small semantic changes: ‘fowl’ and ‘turkey’;
9. Radical semantic changes: ‘broth’ and ‘grease’.

Spelling errors, which may be especially frequent in data that have been acquired through optical character recognition, are easy to detect but have to be corrected manually. Morphological differences (category 2) can be removed by lemmatization. Many of the cases belonging to categories 3 and 4 can be handled by adopting a stop list of determiners, possessive pronouns, and very common modifiers such as *certain, kind of, his, big, female,* etc.

Categories 4, 5, and 6 illustrate a common phenomenon of minor semantic shifts that can be detected without resorting to a lexical resource. All that is needed is the determination of the heads of the phrases, or, more generally, *keywords*. Pairs of glosses that contain matching keywords are usually semantically related.

For the remaining categories, string matching is of no assistance, and some lexical resource is called for. In this paper, I use WordNet (Fellbaum, 1998), or rather, its noun hierarchy, which is the most developed of the four WordNet hierarchies.[3] WordNet is well-suited not only for detecting synonyms but also for associating lexemes that have undergone small 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 favourable sense): ‘woman’ → ‘queen’;
- **pejoration** (developing a less favourable sense): ‘farm-worker’ → ‘villain’;
- **metaphor** (extending the literal meaning): ‘steersman’ → ‘governor’;

[2] For a different “phonetic” approach, based on binary articulatory features, see (Nerbonne and Heeringa, 1997).

[3] The idea of using WordNet for the detection of semantic relationships comes from Lowe and Mazaudon (1994) (footnote 13, page 406).

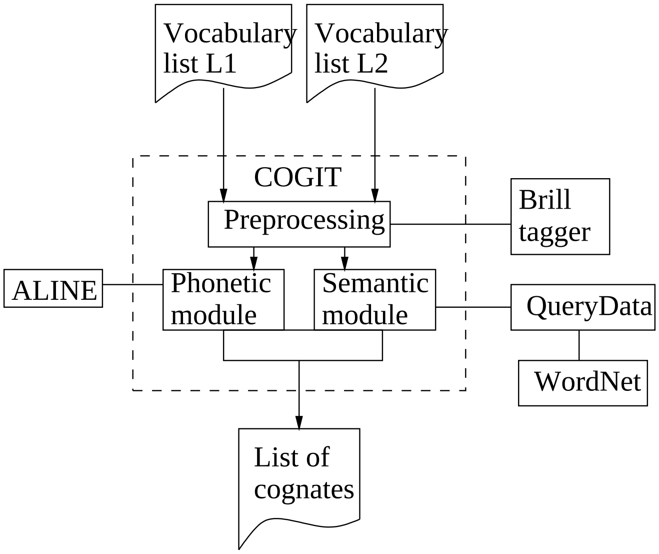

Figure 1: The structure of cognate identification system.

- **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.

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 could be to use a multilingual lexical resource, such as EuroWordNet (Vossen, 1998), which is modeled on the original Princeton WordNet.

5 Implementation

Given two vocabulary lists representing distinct languages, COGIT, the cognate identification system (Figure 1), produces a list of vocabulary-entry pairs, sorted according to the estimated likelihood of their cognateness. Each vocabulary entry consists of a lexeme *l* and its gloss *g*. COGIT is composed of a set of Perl scripts for preprocessing the vocabulary lists, and phonetic and semantic modules written in C++. Both modules return similarity scores in the range $[0,1]$, which are combined into an overall similarity score by the following formula:

$$Sim_{overall}((l_1,g_1),(l_2,g_2)) = (1-\alpha)\cdot Sim_{phon}(l_1,l_2)+\alpha\cdot Sim_{sem}(g_1,g_2),$$

where α is a parameter reflecting the relative importance of the semantic vs. phonetic score. The algorithm is presented informally in Figure 2.

```
1. For each entry in vocabularies $L_1$ and $L_2$:
    (a) Remove stop words.
    (b) Select keywords.
    (c) Perform lemmatization.
    (d) Generate lists of semantically related words.
2. For each pair of entries $(i,j) \in (L_1 \times L_2)$:
    (a) Compute the phonetic similarity score $Sim_{phon}$.
    (b) Compute the semantic similarity score $Sim_{sem}$.
    (c) Set $Sim_{overall} \leftarrow (1-\alpha)\cdot Sim_{phon} + \alpha\cdot Sim_{sem}$.
    (d) If $Sim_{overall} \geq T$, record $i$, $j$, and $Sim_{overall}$.
3. Sort the pairs in descending order of $Sim_{overall}$.
```

Figure 2: Cognate identification algorithm.

The preprocessing of the glosses involves stop word removal and keyword selection. A simple heuristic is used for the latter: the preprocessing script marks as keywords all nouns apart from those that follow a wh-word or a preposition other than “of”. Nouns are identified by a part-of-speech tagger (Brill, 1995), which is applied to glosses after prepending them with the string “It is a”. Checking and correcting the spelling of glosses is assumed to have been done beforehand.

The phonetic module calculates phonetic similarity using either ALINE or a straightforward method such as LCSR, DICE, or truncation. The truncation coefficient is obtained by dividing the length of the common prefix by the average of the lengths of the two words being compared. The similarity score returned by ALINE is also normalized, so that it falls in the range $[0,1]$. The implementation of ALINE is described in (Kondrak, 2000).

For the calculation of a WordNet-based semantic similarity score, I initially used the length of the shortest path between synsets, measured in the num-

Rank	Similarity level	Score	Coverage
1	gloss identity	1.00	.618
2	gloss synonymy	0.70	.020
3	keyword identity	0.50	.105
4	gloss hyponymy	0.50	.023
5	keyword synonymy	0.35	.012
6	keyword hyponymy	0.25	.021
7	gloss meronymy	0.10	.002
8	keyword meronymy	0.05	.000
9	none detected	0.00	.199

Table 3: Semantic similarity levels.

ber of IS-A links.[4] However, I found the effect of considering paths longer than one link to be negligible. Moreover, the process of determining the link distances between all possible pairs of glosses, separately for each pair, was too time-consuming.

Currently, the semantic score is computed by a faster method that employs QueryData, a Perl WordNet[5] module (Rennie, 1999). A list of synonyms, hyponyms, and meronyms is generated for each gloss and keyword in the preprocessing phase. During the execution of the program, regular string matching is performed directly on the listed senses. Words are considered to be related if there is a relationship link between any of their senses. The semantic score is determined according to a 9-point scale of semantic similarity, which is shown in Table 3. The levels of similarity are considered in order, starting with gloss identity. The exact scores corresponding to each level were established empirically. The coverage figures are discussed in Section 6.

The QueryData module also carries out the lemmatization process.

6 Evaluation

COGIT was evaluated on noun vocabularies of four Algonquian languages. The source of the data was machine-readable vocabulary lists that had been used to produce a computer-generated Algonquian dictionary (Hewson, 1993). No grapheme-to-phoneme conversion was required, as the Algonquian lexemes are given in a phonemic transcription. The lists can be characterized as noisy data; they contain many errors, inconsistencies, duplicates, and lacunae. As much as possible, the entries were cross-checked with the dictionary itself, which is much more consistent. The dictionary, which contains entries from the four languages grouped in cognate sets, also served as a reliable source of cognateness information. Table 4 specifies the number of lexemes available for each language. Only about a third of those nouns are actually in the dictionary; the rest occur only in the vocabulary lists. Table 5 shows the number of cognate pairs for each language combination. To take the Menomini–Ojibwa pair as an example, the task of the system was to identify 259 cognate-pairs from 1540 × 1023 possible lexeme-pairs. The average ratio of non-cognate to cognate pairs was about 6500.

Cree (Cr)	1628
Fox (Fx)	575
Menomini (Mn)	1540
Ojibwa (Oj)	1023

Table 4: Number of lexemes available for each language.

	Cr	Fx	Mn	Oj
Cr	-	130	239	408
Fx	130	-	121	136
Mn	239	121	-	259
Oj	408	136	259	-

Table 5: Number of shared cognates.

Experimental results support the intuition that both the phonetic and the semantic similarity between cognates is greater than between randomly selected lexemes. Table 6 contrasts phonetic similarity scores for cognate pairs and for randomly selected pairs, averaged over all six combinations of languages. The average value of the semantic similarity score, as defined in Table 3, was .713 for cognate pairs, and less than .003 for randomly selected pairs.

	Cognate		Random	
	$\bar{x}$	s	$\bar{x}$	s
Truncation	.284	.267	.012	.041
DICE	.420	.246	.062	.090
LCSR	.629	.155	.236	.101
ALINE	.627	.135	.218	.083

Table 6: Average phonetic similarity between cognate pairs and between randomly selected pairs. $\bar{x}$ - mean; s - standard deviation.

[4] A number of more sophisticated methods exist for measuring semantic similarity using WordNet (Budanitsky, 1999).

[5] The version of WordNet used is 1.6.

	Development	Test Sets	
	Set	$\bar{x}$	s
Truncation	.142	.055	.056
DICE	.411	.078	.086
LCSR	.614	.189	.117
ALINE	.633	.393	.076
Method G	.811	.616	.049
Method K	.826	.642	.052
Method W	.829	.657	.057

Table 7: Interpolated 3-point average precision of various cognate indentification methods. Methods G, K, and W use ALINE combined with increasingly complex semantic similarity detection ($\alpha = 0.2$).

The values of all parameters, including α, ALINE's parameters[6], and the semantic similarity scale given in Table 3, were established during the development phase of the system, using only the Cree–Ojibwa data. These two languages were chosen as the development set because they are represented by the most complete vocabularies and share the largest number of cognates. However, as it turned out later, they are also the most closely related among the four Algonquian languages, according to all measures of phonetic similarity. It is quite possible that the overall performance of the system would have been better if a different language pair had been chosen as the development set.

Table 7 compares the effectiveness of various cognate identification methods, using interpolated 3-point average precision. The first four methods (Truncation, DICE, LCSR, and ALINE) are based solely on phonetic similarity. The remaining three methods combine ALINE with increasingly sophisticated semantic similarity detection: Method G considers gloss identity only, Method K adds keyword-matching, and Method W employs also WordNet relations. The results for the development set (Cree–Ojibwa) are given in the first column. The results for the remaining five sets are given jointly as their average and standard deviation.

The choice of 3-point average precision requires explanation. The output of the system is a sorted list of suspected cognate pairs. 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 for the levels 20%, 50%, and 80%, and then averaged to yield a single number.

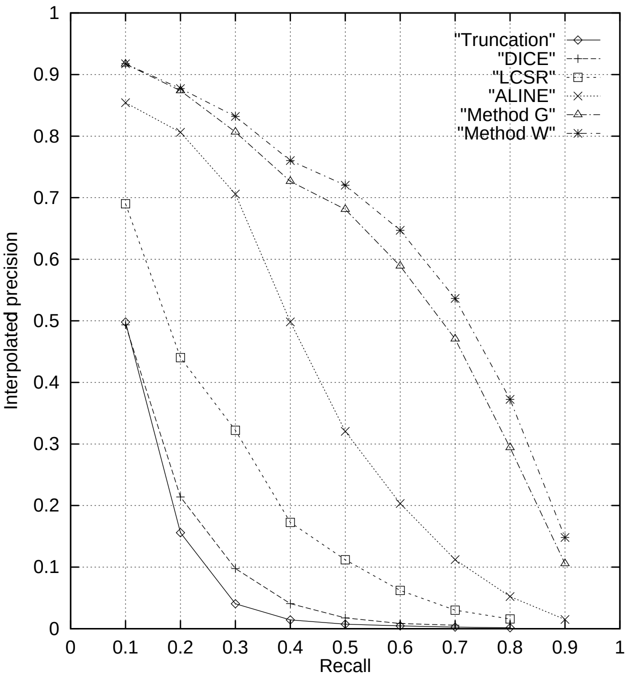

Figure 3: Precision-recall curves for various methods.

Figure 3 shows a more detailed comparison of the effectiveness of the methods on test sets, in the form of precision–recall curves. Among the phonetic methods, ALINE outperforms all "orthographic" coefficients, including LCSR, The dominance of ALINE increases as more remote languages are considered. Dice's coefficient performs poorly as a cognate identification method, being only slightly better than a naive truncation method. All three methods that use the semantic information provided by the glosses perform substantially better than the purely phonetic methods. Impressive results are reached even when only gloss identity is considered. Adding keyword-matching and WordNet relations brings additional, albeit modest, improvements.[7] When, instead of ALINE, LCSR is used in conjunction with the semantic methods, the

[6] ALINE's parameters were set as follows: C_{skip} = –1, C_{sub} = 10, C_{exp} = 15 and C_{vwl} = 1. The salience settings were the same as in (Kondrak, 2000), except that the salience of feature "Long" was set to 5.

[7] The curve for Method K, which would be slightly below the curve for Method W, is omitted for clarity.

average precision numbers are lower by over 10 percentage points.

Figure 4 illustrates the effect of varying the setting of the parameter α on the average precision of COGIT when ALINE is used in conjunction with full semantic analysis. The greater the value of α, the more weight is given to the semantic score, so $\alpha = 0$ implies that the semantic information is ignored. The optimal value of α for both the development and the test sets is close to 0.2. With α approaching 1, the role of the phonetic score is increasingly limited to ordering candidate pairs within semantic similarity levels. Average precision plummets to 0.161 when α is set to 1 and hence no phonetic score is available.

The rightmost column in Table 3 in Section 5 compares proportions of all cognate pairs in the data that are covered by individual semantic similarity levels. Over 60% of cognates have at least one gloss in common. (However, only about one in four pairs sharing a gloss are actual cognates.) The cases in which the existence of a WordNet relation influences the value of the similarity score account for less than 10% of the cognate pairs. In particular, instances of meronymy between cognates are very rare.

Apart from the limited coverage of WordNet-related semantic similarity levels, there are other reasons for the relatively small contribution of WordNet to the overall performance of the system. First, even after preprocessing that includes checking the spelling, lemmatization, and stop word removal, many of the glosses are not in a form that can be recognized by WordNet. These include compounds written as a single word (e.g. ‘snowshoe’), and rare words (e.g. ‘spawner’) that are not in WordNet. Second, when many words have several meanings that participate in different synsets, the senses detected to be related are not necessarily the senses used in the glosses. For example, ‘star’ and ‘lead’ share a synset (“an actor who plays a principal role”), but in the Algonquian vocabularies both words are always used in their most literal sense. Only in the case of complete identity of glosses can the lexemes be assumed to be synonymous in all senses. Finally, since the data for all Algonquian languages originates from a single project, it is quite homogeneous. As a result, many glosses match perfectly within cognate sets, which limits the need for application of WordNet lexical relations.

The performance figures are adversely affected by the presence of the usual “noise”, which is unavoidable in the case of authentic data. Manual preparation of the vocabulary lists would undoubtedly result in better performance. However, because of its size, only limited automatic validation of the data had been performed. It should also be noted that examination of apparent false positives sometimes leads to discovering true cognates that are not identified as such in Hewson’s dictionary. One interesting example is Cree *pīsākanāpiy* ‘rope, rawhide thong’, and Ojibwa *pīššākaniyāp* ‘string’. In this case COGIT detected the synonymy of the glosses by consulting WordNet.

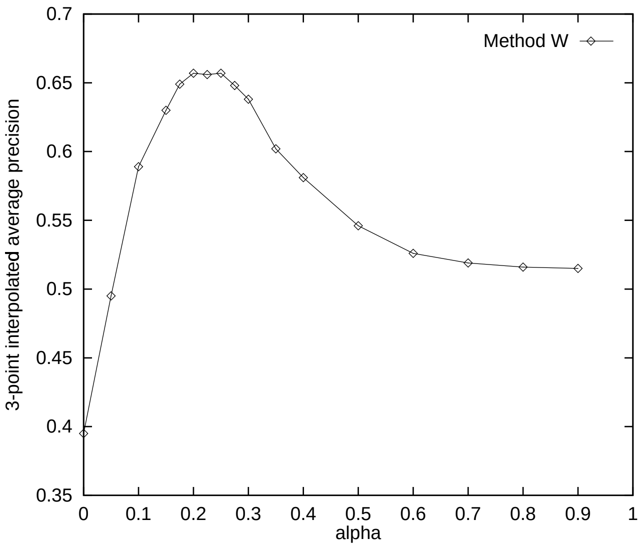

Figure 4: Interpolated 3-point average precision of Method W on test sets as a function of the parameter α, which reflects the relative importance of the semantic vs. phonetic similarity.

7 Conclusion

The results show that it is possible to identify a large portion of cognates in related languages without explicit knowledge of systematic sound correspondences between them or phonological changes that they have undergone. This is because cognates on average display higher phonetic and semantic similarity than words that are unrelated. Many vocabulary entries can be classified as cognates solely on the basis of their phonetic similarity. ALINE, a sophisticated algorithm based on phonological features, is more successful at this task than simple “orthographic” measures. Analysis of semantic information extracted from glosses yields a dramatic increase in the number of identified cognates. Most of the improvement comes from detecting entries that have matching glosses. On the other hand, the contribution of WordNet is small.

A system such as COGIT can be of assistance for comparative linguists dealing with large vocabulary data from languages with which they are unfamiliar. It can also serve as one of the principal modules of a language reconstruction system. However, in spite of the fact that the main focus of this paper is diachronic phonology, the techniques and findings presented here may also be applicable in other contexts where it is necessary to identify cognates, such as bitext alignment.

Acknowledgments

Thanks to Graeme Hirst, Elan Dresher, Suzanne Stevenson, Radford Neal, and Gerald Penn for their comments, to John Hewson for the Algonquian data, and to Alexander Budanitsky for the semantic distance code. This research was supported by Natural Sciences and Engineering Research Council of Canada.

References

Chris Brew and David McKelvie. 1996. Word-pair extraction for lexicography. In K. Oflazer and H. Somers, editors, *Proceedings of the Second International Conference on New Methods in Language Processing*, pages 45–55, Ankara, Bilkent University.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–566.

Alexander Budanitsky. 1999. Lexical semantic relatedness and its application in natural language processing. Technical Report CSRG-390, University of Toronto. Available from ftp.cs.toronto.edu/csrg-technical-reports.

Kenneth W. Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Columbus, Ohio.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. The MIT Press, Cambridge, Massachusetts.

Jacques B. M. Guy. 1994. An algorithm for identifying cognates in bilingual wordlists and its applicability to machine translation. *Journal of Quantitative Linguistics*, 1(1):35–42.

John Hewson. 1974. Comparative reconstruction on the computer. In *Proceedings of the First International Conference on Historical Linguistics*, pages 191–197.

John Hewson. 1993. *A computer-generated dictionary of proto-Algonquian*. Hull, Quebec: Canadian Museum of Civilization.

John Hewson. 1999. Vocabularies of Fox, Cree, Menomini, and Ojibwa. Computer file.

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

John B. Lowe and Martine Mazaudon. 1994. The reconstruction engine: a computer implementation of the comparative method. *Computational Linguistics*, 20:381–417.

Tony McEnery and Michael Oakes. 1996. Sentence and word alignment in the CRATER Project. In J. Thomas and M. Short, editors, *Using Corpora for Language Research*, pages 211–231. Longman.

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

John Nerbonne and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-97)*. Available from http://www.cogsci.ed.ac.uk/sigphon.

Jason Rennie. 1999. Wordnet::QueryData Perl module. Available from http://www.ai.mit.edu/˜jrennie.

Don Ringe. 1998. A probabilistic evaluation of Indo-Uralic. In Joseph C. Salmons and Brian D. Joseph, editors, *Nostratic: sifting the evidence*, pages 153–197. Amsterdam: John Benjamins.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. 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.

R. L. Trask. 1996. *Historical Linguistics*. London: Arnold.

Piek Vossen, editor. 1998. *EuroWordNet: a Multilingual Database with Lexical Semantic Networks*. Kluwer Academic, Dordrecht.