Identifying Complex Sound Correspondences in Bilingual Wordlists

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. The determination of recurrent sound correspondences between languages is crucial for the identification of cognates, which are often employed in statistical machine translation for sentence and word alignment. In this paper, an algorithm designed for extracting non-compositional compounds from bitexts is shown to be capable of determining complex sound correspondences in bilingual wordlists. In experimental evaluation, a C++ implementation of the algorithm achieves approximately 90% recall *and* precision on authentic language data.

1 Introduction

All languages change through time. Table 1 gives an example of how much English has evolved within the last fourteen hundred years. Words that make up languages undergo sound changes (*nū* → *now*) as well as semantic shifts ('guardian' → 'ward'). Lexical replacement is a process in which lexemes drop out of usage altogether, and are substituted by other, unrelated words (*herigean* → *praise*). Morphological endings change and disappear as well (*-on* in *sculon*).

Old English:	Nū	sculon	herigean	heofonrīces	weard
Modern English:	Now	we should	praise	heaven-kingdom's	guardian

Table 1. The first verse of Caedmon's *Hymn* and its modern English translation.

When two groups of people that speak a common language lose contact with each other, their respective languages begin to diverge, and eventually become mutually unintelligible. In such cases, we may still be able to determine that the languages are genetically related by examining cognates, that is words that have developed from the same proto-form. For example, French *lait*, Spanish *leche*, and Italian *latte* constitute a cognate set, as they are all descendants, or reflexes, of Latin *lacte*. In general, the longer the time that has passed since the linguistic

split, the smaller the number of cognates that remain as a proof of a genetic relationship.

Because of gradual changes over long periods of time, cognates often acquire very different phonetic shapes. For example, English *hundred*, French *cent*, and Polish *sto* are all descendants of Proto-Indo-European **kmtom* (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, not all similar sounding words that have the same meaning are cognates. It can be a matter of chance resemblance, as in English *day* and Latin *die* 'day', or an instance of a borrowing, as in English *sprint* and Japanese *supurinto*. Borrowings are lexical items that have been incorporated (possibly in modified form) into one language from another.

An important phenomenon that allows us to distinguish between cognates and borrowings is the regularity of sound change. The regularity principle states that a change in pronunciation applies to sounds in a given phonological context across all words in the language. Regular sound changes tend to produce regular correspondences of phonemes in corresponding cognates. /d/:/t/ is a regular correspondence between English and German, as evidenced by cognate pairs such as *day – tag*, *dry – trocken*, and *drink – trinken*. Table 2 shows contains examples of a regular sound correspondence between four Romance languages. I prefer to use the term *recurrent sound correspondences* because in practice the matchings of phonemes in cognate pairs are more tendencies than hard-and-fast rules.

Latin	*Italian*	*Spanish*	*French*	
no**ct**e	no**tt**e	no**ch**e	nu**it**	'night'
o**ct**o	o**tt**o	o**ch**o	hu**it**	'eight'
la**ct**e	la**tt**e	le**ch**e	la**it**	'milk'
fa**ct**u	fa**tt**o	he**ch**o	fa**it**	'done'
te**ct**u	te**tt**o	te**ch**o	to**it**	'roof'

Table 2. An example of a recurrent sound correspondence in related languages.

The determination of recurrent sound correspondences is the principal step of the comparative method of language reconstruction. Not only does it provide evidence for the relatedness of languages, but it also makes it possible to distinguish cognates from borrowings and chance resemblances. However, because manual determination of recurrent sound correspondences is an extremely time-consuming process, it has yet to be accomplished for many proposed language families. A system able to perform this task automatically from unprocessed bilingual wordlists could be of great assistance to historical linguists. The *Reconstruction Engine* [14], a set of programs designed to be an aid in language

reconstruction, requires a set of recurrent sound correspondences to be provided beforehand.

The determination of recurrent sound correspondences is closely related to another task that has been much studied in computational linguistics, the identification of cognates. Cognates have been employed for sentence and word alignment in bitexts [16], improving statistical machine translation models [1], and inducing translation lexicons [10]. Some of the proposed cognate identification algorithms implicitly determine and employ recurrent sound correspondences [18, 15].

Although it may not be immediately apparent, there is 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. The consistency with which a word in one language is translated into a word in another language is mirrored by the consistency of sound correspondences. The former is due to the semantic relation of synonymy, while the latter follows from the principle of the regularity of sound change. Thus, as already asserted by Guy [5], it should be possible to use similar techniques for both tasks.

The method of determining complex recurrent sound correspondences that I present here adopts the approach proposed in [13]. The idea is to relate correspondences between sounds in wordlists to translational equivalences between words in bitexts (bilingual corpora). The method induces models of sound correspondence that are similar to models developed for statistical machine translation. It has been shown [13] that the method is able to determine recurrent sound correspondences with high accuracy in bilingual wordlists in which less than 30% of the pairs are cognates. However, in the one-to-one model employed by the method, links are induced only between individual phonemes. This is a serious limitation because recurrent sound correspondences often involve clusters of phonemes. Many-to-many correspondences, such as the ones shown in Table 2, may either be only partially recovered or even completely missed by the algorithm.

This paper presents an extension of the approach described in [13], which overcomes its main limitation by adapting the algorithm for discovering *non-compositional compounds* (NCCs) in bitexts proposed by Melamed [16]. In Section 2, I review previous work on determination of recurrent sound correspondences. Melamed's approach to inducing models of translational equivalence is discussed in Section 3. Section 4 describes the algorithm for discovering non-compositional compounds. Section 5 contains some implementation details. Section 6 describes the data used for the experimental evaluation, and Section 7 is devoted to the evaluation itself.

2 Related Work

In a schematic description of the comparative method, the two steps that precede the determination of recurrent sound correspondences are the identification of cognate pairs [12], and their phonetic alignment [11]. Indeed, if a compre-

hensive set of correctly aligned cognate pairs is available, the recurrent sound correspondences could be extracted by simply following the alignment links. Unfortunately, in order to make reliable judgments of cognation, it is necessary to know in advance what the recurrent sound correspondences are. Historical linguists solve this apparent circularity by guessing a small number of likely cognates and refining the set of correspondences and cognates in an iterative fashion.

Guy [5] outlines an algorithm for identifying cognates in bilingual wordlists which is based on recurrent sound correspondences. The algorithm estimates the probability of phoneme correspondences by employing a variant of the χ^2 statistic on a contingency table, which indicates how often two phonemes co-occur in words of the same meaning. The probabilities are then converted into the estimates of cognation by means of some experimentation-based heuristics. Only simple, one-to-one phoneme correspondences are considered. The paper does not contain any evaluation on authentic language data, but Guy's program COGNATE, which implements the algorithm, is publicly available. The program does not output an explicit list of recurrent sound correspondences, which makes direct comparison with my method difficult.

Oakes [17] describes a set of programs that together perform several steps of the comparative method, from the determination of recurrent sound correspondences in wordlists to the actual reconstruction of the proto-forms. Word pairs are considered cognate if their edit distance is below a certain threshold. The edit operations cover a number of sound-change categories. Sound correspondences are deemed to be regular if they are found to occur more than once in the data. The paper describes experimental results of running the programs on a set of wordlists representing four Indonesian languages, and compares those to the reconstructions found in the linguistic literature. Section 7 contains a comparison of the recurrent sound correspondences identified by JAKARTA and the ones discovered by my method.

Because the tasks of determination of recurrent sound correspondence and the identification of cognates are intertwined, some of the bitext-related algorithms implicitly determine and employ recurrent sound correspondences. Tiedemann [18] considers automatic construction of weighted string similarity measures from bitexts. He includes three lists of the most frequent character "mappings" between Swedish and English, which correspond to his three mapping approaches (single characters, vowel and consonant sequences, and non-matching parts of two strings). However, because genetic cognates in the data seem to be outnumbered by borrowings, the lists contain few genuine correspondences. Mann and Yarowsky [15] take advantage of language relatedness in order to automatically induce translation lexicons. In their search for cognates, they discover most probable character "substitutions" across languages. In the provided French–Portuguese examples, phonologically plausible correspondences *b:v, t:d* mix with mere orthographic regularities *c:q, x:s*.

Knight and Graehl [9] in their paper on back-transliteration from the Japanese syllabic script *katakana* to the English orthography consider the sub-task of

aligning the English and Japanese phonetic strings. They apply the estimation-maximization (EM) algorithm to generate symbol-mapping probabilities from 8,000 pairs of unaligned English–Japanese sound sequences. It is possible to view the sound pairs with the highest probabilities as the strongest recurrent correspondences between the two languages. Naturally, the existence of those correspondences is an artifact of the transliteration process, rather than a consequence of a genetic language relationship. Nevertheless, it may be possible to employ a similar approach to discover recurrent sound correspondences in genuine cognates. A drawback of the alignment model presented in the paper is an asymmetric, one-to-many mapping between the English and Japanese sounds, and a restricted set of edit operations that excludes both insertions and deletions. These restrictions are designed to make the models less expensive to compute.

3 The Word-to-Word Model of Translational Equivalence

In statistical machine translation, 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, or alignment, of words between two parts of a bitext is not known in advance.

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

Melamed [16] shows that the assumption of independence leads to invalid word associations, and proposes an algorithm for inducing models of translational equivalence that outperform 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 scores are initialized according to the G^2 statistic [4]. 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. Three translation-model re-estimation methods are possible: Method A calculates the likelihood scores as the logarithm of the probability of jointly generating the pair of words, Method B uses auxiliary parameters to represent an explicit noise model, and Method C conditions the auxiliary pa-

rameters on various word classes. The re-estimated likelihood scores are then applied to find a new set of links. The process is repeated until the translation model converges to the desired degree.

As demonstrated in [13], it is possible to adapt Melamed's algorithm to the problem of determining recurrent sound correspondences. The main idea is to induce a model of sound correspondence in a bilingual wordlist, in the same way as 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 recurrent sound correspondences.

The most important modification to the original algorithm is the substitution of the approximate competitive-linking algorithm of Melamed with a variant of the well-known dynamic programming algorithm [11], which computes the *optimal* alignment between two strings in polynomial time. Insertion and deletion of segments is modeled by employing an *indel* penalty for unlinked segments, rather than by *null links* used by Melamed. The alignment score between two words is computed by summing the number of induced links, and applying an indel penalty for each unlinked segment, with the exception of the segments beyond the rightmost link. In order to avoid inducing links that are unlikely to represent recurrent sound correspondences, only pairs whose likelihood scores exceed a set threshold are linked.

The algorithm for the determination of recurrent sound correspondences was evaluated on 200-word lists of basic meanings representing several Indo-European languages. The results show that the method is capable of determining recurrent sound correspondences in bilingual wordlists in which less than 30% of pairs are cognates, and that it outperforms comparable algorithms on the related task of the identification of cognates.

4 Discovering Non-Compositional Compounds in Bitexts

The algorithm proposed in [13] can only discover recurrent sound correspondences between single phonemes. This limitation, which is directly inherited from Melamed's original algorithm, may prevent the algorithm from detecting many more complex correspondences, such as the ones in Table 2. 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.

As a way of relaxing the *one-to-one* restriction, Melamed [16] 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 objec-

tive function that measures the predictive power of a translation model $Pr(s,t)$, Melamed selects *mutual information*:

$$I(S;T) = \sum_{s \in S} \sum_{t \in T} Pr(s,t) \log \frac{Pr(s,t)}{Pr(s)Pr(t)},$$

where S and T represent the distributions of linked words in the source and target texts, and s and t are word tokens.

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. The contribution of s to $I(S;T)$ is given as:

$$i(s) = \sum_{t \in T} Pr(s,t) \log \frac{Pr(s,t)}{Pr(s)Pr(t)}.$$

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. Each candidate NCC xy causes the net change Δ_{xy} in the objective function, which can be expressed as:

$$\Delta_{xy} = i'(x) + i'(y) + i'(xy) - i(x) - i(y),$$

where i and i' are predictive value functions for source words in the base translation model and in the trial translation model, respectively. $i'(x)$ is estimated on the assumption that the links involving x will not change in the trial translation model unless y occurs to the right of x:

$$i'(x) = i(x : RC \neq y),$$

where $(x : RC \neq y)$ denotes the set of tokens of x whose right context is y. Similarly,

$$i'(y) = i(y : LC \neq x),$$

where LC denotes word context to the left. Finally, $i'(xy)$ is estimated as follows:

$$i'(xy) = i(x : RC = y) + i(y : LC = x).$$

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 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 reaching an application-dependent stopping condition.

The method was evaluated on a large English–French bitext containing transcripts of Canadian parliamentary debates (Hansards). In one experiment, after six iterations the algorithm identified on both sides of the bitext about four hundred NCCs that increased the mutual information of the model. Another experiment, which is particularly relevant for the application discussed in this chapter, showed that the method was capable of discovering meaningful NCCs in a data set consisting of spellings and pronunciations of English words (for example, *ph* was determined to be a NCC of English spelling because it consistently "translates" into the sound /f/). However, the full NCC recognition algorithm was not tested in any real application.

5 Implementation of the Algorithm

The NCC algorithm of Melamed has been adapted to the problem of determining complex sound correspondences and implemented as a C++ program named CORDI. The program takes as input a bilingual wordlist and produces an ordered list of recurrent sound correspondences. Method C discussed in Section 3 is used for the inducing of translation models, 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.

Adjustable parameters include the indel penalty ratio d and the minimum-strength correspondence threshold t. The parameter d controls the behaviour of the alignment algorithm by fixing the ratio between the negative indel weight and the positive weight assigned to every induced link. A lower ratio causes the program to be more adventurous in positing sparse links. The parameter t controls the tradeoff between reliability and the number of links. The value of t implies a score threshold of $t \cdot \log \frac{\lambda^+}{\lambda^-}$, which is a score achieved by a pair of phonemes that have t links out of t co-occurrences. In all experiments described below, d was set to 0.15, and t was set to 1 (sufficient to reject all non-recurring correspondences). The maximum number of iterations of the NCC algorithm should also be specified by the user, but the algorithm may terminate sooner if two subsequent iterations fail to produce any candidate NCCs.

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, I decided to accept all candidate NCCs that result in a recurrent sound correspondence that has a likelihood score above the minimum-strength threshold t described above. I found that the strength of an induced correspondence better reflects the importance of a phoneme cluster than the mutual information gain criterion.

6 The Algonquian Data

The test data suitable for the evaluation of the approach outlined above has to fulfill several requirements: it should be sufficiently large to contain many surviving cognates, the lexemes should be given in a consistent notation that allows for an automatic transcription into phonetic form, and, finally, the cognation information has to be provided in the electronic form as well, so that the performance of the program can be measured objectively. The last condition is perhaps the most difficult to satisfy. Even in the rare cases when machine-readable bilingual lexicons can be acquired, the cognation judgments would have to be laboriously extracted from etymological dictionaries. Note that optical scanning of phonetic symbols or unusual diacritics is is not feasible with the current state of technology.

Fortunately, the machine-readable Algonquian data [8] satisfy the above requirements. It consists of two parts that complement each other: the etymological dictionary, and the vocabulary lists from which the dictionary was produced.

The dictionary, which is also available in book form [7], contains 4,068 cognate sets, including 853 marked as nouns. Each cognate set is composed of a reconstructed proto-form and the corresponding cognates accompanied by short glosses in English. Nearly all cognates belong to one of the four principal Algonquian languages (Fox, Menomini, Cree, Ojibwa). The dictionary file is almost identical with the book version, and required only minimal clean-up. The lexemes are already in a phonemic transcription, so no sophisticated grapheme-to-phoneme conversion was necessary. A simple coding is used to express phonemes that lack ASCII equivalents: *c* for /š/, *q* for the glottal stop, etc. In the experiments described in this section, the dictionary file served as a source of the cognation information.

Language	Dictionary only		Dictionary and lists	
	All words	Nouns	All words	Nouns
Fox	1252	193	4759	575
Menomini	2231	361	8550	1540
Cree	2541	512	7507	1628
Ojibwa	2758	535	6334	1023
Total	**8782**	**1601**	**27150**	**4766**

Table 3. The size of the Algonquian vocabulary lists.

In contrast with the dictionary, the vocabulary lists can be characterized as noisy data. They contain many errors, inconsistencies, duplicates, and lacunae. The Fox file is incomplete. In the Menomini file, three different phonemes (/č/, /æ/, and the glottal stop) had been merged into one, and had to be painstakingly reconstructed on the basis of phonotactic constraints. As much as possible, the entries were cross-checked with the dictionary itself, which is much more consis-

tent. Table 3 specifies the number of unique lexemes available for each language. It appears that only about a third of the nouns present in the vocabulary lists had made it into the dictionary.

7 Experimental Evaluation

In order to test the suitability of the NCC approach, an experiment was performed on a subset of the Algonquian data. The goal was to determine recurrent sound correspondences from noisy wordlists and evaluate them against the set of correspondences determined by Bloomfield [2, 3]. Because of the large number of complex 1:2 and 2:2 recurrent sound correspondences, the Algonquian languages are ideal for testing the NCC approach.

The input data was automatically extracted from the raw vocabulary lists by selecting all pairs of noun lexemes that had at least one gloss in common. The end result of such an operation is bilingual wordlists containing both cognate and non-cognate pairs. The Cree–Ojibwa list served as the development set, and the Fox–Menomini list as the test set. The Cree–Ojibwa contained 732 pairs, including 242 (33.1%) cognate pairs. The Fox–Menomini list turned out to be even more challenging: it contained 397 word pairs, including only 79 (19.9%) cognate pairs.

Since the vowel correspondences in Algonquian are rather inconsistent, following Hewson [6], I decided to concentrate on consonants and consonant clusters. 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 4 compares the set of 31 correspondences enumerated by Bloomfield, which is adopted as the gold standard, with the set of 23 correspondences determined by CORDI, and eight correspondences identified by JAKARTA [17]. 20 recurrent sound correspondences identified by CORDI are correct, while the remaining three are wrong and can be traced to alignments of unrelated words. The resulting precision was therefore 87%.

Bloomfield:	**p:p t:t k:k s:s h:h** h:q **č:č š:s n:n m:m t:ht hp:hp hk:hk ht:qt hk:hk šk:sk č:hč s:hs s:qs š:qs s:hn s:qn šk:hk p:hp hč:qč** k:hk hk:čk hp:sp hč:hč ht:ht š:hs n:hn
CORDI:	**p:p t:t k:k s:s h:h č:č š:s** p:č **n:n m:m t:ht hp:hp hk:hk ht:qt hk:hk šk:sk č:hč s:hs s:qs š:qs s:hn s:qn** hk:t t:sk
JAKARTA:	**p:p t:t k:k s:s h:h n:n m:m** h:hs

Table 4. The Fox–Menomini consonantal correspondences determined by a linguist and by two computer programs. The correspondences shown in boldface are valid correspondences that were present in the input set of word pairs.

In order to determine why the number of recurrent sound correspondences established by Bloomfield was much greater than the number of recurrent sound correspondences produced by the program, I manually analyzed the 79 cognate pairs included in the input wordlist. I found that *š:hk* and *p:hp* occur twice in the input, *hč:qč* occurs once, and the remaining seven complex correspondences do not occur at all. The *h:q* correspondence is dubious because it only occurs within clusters. Since, by definition, recurrent correspondences are those that occur at least twice, the recall on the test set was in fact $21/23 = 91\%$.

For comparison, on the same Fox–Menomini list, JAKARTA identifies only eight consonantal correspondences of which the single complex correspondence is not in Bloomfield's set. The resulting precision is comparable at 88%, but the recall is only 32%.

The results of the experiment are extremely encouraging. The accomplishment of a very high precision *and* recall on a test set composed of 80% noise confirms that the iterative statistical approach advocated here is highly robust. The impressive outcome should, however, be interpreted with caution. Because of the (unavoidably) small number of target correspondences, the change of a single classification makes a difference of about 5% in the resulting precision/recall figures. Moreover, the decision to ignore vowels and glides helped the program to focus on the right type of correspondences. Finally, the Algonquian consonantal correspondences are almost context-free, which nicely suits the program's principles.

8 Conclusion

I have proposed an original approach to the determination of complex sound correspondences in bilingual wordlists based on the idea of relating recurrent correspondences between sounds to translational equivalences between words. Through induction of statistical models that are similar to those developed for statistical machine translation, the method is able to recover recurrent sound correspondences from bilingual wordlists that consist mostly of unrelated pairs. The results presented here prove that the techniques developed in the context of statistical machine translation can be successfully applied to a problem in diachronic phonology. I am convinced that the transfer of methods and insights is also possible in the other direction.

Acknowledgments

I would like to thank Graeme Hirst, Radford Neal, and Suzanne Stevenson for helpful comments, to John Hewson for the Algonquian data, to Michael Oakes for assistance with JAKARTA, 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. Leonard Bloomfield. On the sound-system of central Algonquian. *Language*, 1:130–156, 1925.
3. Leonard Bloomfield. Algonquian. In Harry Hoijer et al., editor, *Linguistic Structures of Native America*, volume 6 of *Viking Fund Publications in Anthropology*, pages 85–129. New York: Viking, 1946.
4. Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
5. 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.
6. John Hewson. Comparative reconstruction on the computer. In *Proceedings of the 1st International Conference on Historical Linguistics*, pages 191–197, 1974.
7. John Hewson. *A computer-generated dictionary of proto-Algonquian*. Hull, Quebec: Canadian Museum of Civilization, 1993.
8. John Hewson. Vocabularies of Fox, Cree, Menomini, and Ojibwa, 1999. Computer file.
9. Kevin Knight and Jonathan Graehl. Machine transliteration. *Computational Linguistics*, 24(4):599–612, 1998.
10. 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.
11. 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.
12. 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.
13. 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.
14. John B. Lowe and Martine Mazaudon. The reconstruction engine: a computer implementation of the comparative method. *Computational Linguistics*, 20:381–417, 1994.
15. 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.
16. I. Dan Melamed. *Empirical Methods for Exploiting Parallel Texts*. The MIT Press, Cambridge, MA, 2001.
17. 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.
18. 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.