

Estimating the Resource Adaption Cost from a Resource Rich Language to a Similar Resource Poor Language

Anil Kumar Singh, Kiran Pala and Harshit Surana

Language Technologies Research Centre
IIIT, Hyderabad, India
anil@research.iiit.ac.in, kiranap@gmail.com, surana.h@gmail.com

Abstract

Developing resources which can be used for Natural Language Processing is an extremely difficult task for any language, but is even more so for less privileged (or less computerized) languages. One way to overcome this difficulty is to adapt the resources of a linguistically close resource rich language. In this paper we discuss how the cost of such adaption can be estimated using subjective and objective measures of linguistic similarity for allocating financial resources, time, manpower etc. Since this is the first work of its kind, the method described in this paper should be seen as only a preliminary method, indicative of how better methods can be developed. Corpora of several less computerized languages had to be collected for the work described in the paper, which was difficult because for many of these varieties there is not much electronic data available. Even if it is, it is in non-standard encodings, which means that we had to build encoding converters for these varieties. The varieties we have focused on are some of the varieties spoken in the South Asian region.

1. Introduction

In a linguistically dense and diverse area like South Asia, the number of languages with a large number of speakers is quite high. At the same time, the resources for these languages are very scarce, and so are other resources (finance, time, manpower etc.) which are required for creating the language resources. However, there is one fact which can make the task of building resources for many of these languages somewhat easier. This fact is the similarity of languages (or varieties) belonging to certain groups or families, which is partly the result of common origin and partly of long historical proximity (Emeneau, 1956; Emeneau, 1980). Even the varieties belonging to different language families can be very similar at certain linguistic levels.

In such a situation, the ability to quantitatively measure the similarities and differences among languages as well as dialects (Dyen et al., 1992; Nerbonne and Heeringa, 1997; Kondrak, 2002; Ellison and Kirby, 2006) can be very important, not only for providing evidence for or against a variety being a language or a dialect, but also for the more practical purpose of building resources for all these varieties, especially those which are relatively less privileged.

When two languages (or varieties) are quite close and one of them happens to be a more privileged one in the sense of having language resources, it may be possible to adapt the resources of the more privileged variety for the less privileged one with much less effort than would be required if those resources were to be created from scratch. Of course, not all resources can be adaptable, even for close languages. Still, in the South Asian context, it is quite likely that a more privileged variety will be close to more than one less privileged variety. In such a case, the more privileged variety can be treated as a ‘source’ language or the variety around which the development of the resources for a number of ‘target’ varieties can be centered. Here the terms ‘source’ and ‘target’ are being used to suggest that the task of resource adaption can be seen (somewhat metaphorically) as

a kind of resource ‘translation’. As an example, Hindi can be the source variety for Braj, Rajasthani, Avadhi, Bhojpuri (which are *considered* to be the dialects of Hindi) and perhaps even for Punjabi (which is *considered* to be a separate language).

In this paper, we present the results of some experiments in trying to estimate the cost of developing resources for less privileged languages provided that there is a source language with some existing resources. We take some such source and some less privileged languages and try to estimate the cost in terms of different kinds of linguistic distances. We also discuss how such estimation of cost can be performed in a more subjective way, using the knowledge about the linguistic characteristics of the languages being considered. Since this is a first work of its kind, the method for estimation described here should be seen as only a preliminary method. Hopefully, in future, much better ways of estimation will be developed. One purpose for which such estimates of the cost of resource adaption can be used is allocation of resources for building resources through partial or complete adaption.

2. Similarities of Languages

The way we study the differences and the similarities among languages depends on our purpose. So, for example, if we want to construct a genealogy of the world’s languages (Nakleh et al., 2005; Nerbonne, 2005), we might be able to achieve our goals even if we restrict our study to phonology and lexicon. On the other hand, the study of typology and universals (McMahon and McMahon, 2005) will require us to cover other linguistic levels such as syntax and semantics. Similarly, when we want to estimate the cost of adapting resources of one language for another, our method for estimation will depend on the kind of resource we are trying to adapt. Strictly speaking, this is true. However, there is one interesting question: do the differences among languages at one level (say, phonology) roughly also give an estimate of difference at another level (say, syntax)? The argument could be that if two languages are genetically

distant (as found by studies at the levels of phonology and lexicon), they are also likely to be distant at the syntactic level because, in such a case, syntactic similarities can only be accidental. This argument implies that it is reasonable to assume that the distance among languages calculated on the basis of study at one or two levels roughly generalizes to other levels too. However, there can be exceptions. For example, Italian and Japanese are phonologically similar but are syntactically very different. But it should be noted that here we are only interested in the similarities (or differences) among varieties which we do know to be similar. The problem is to get a quantitative estimate of the similarity (or the distance) for the purposes of resource adaption.

3. Cost of Resource Adaption

It is obvious that not all resources can be adapted even for very close languages. For example, Hindi and Urdu are close enough to be considered the same language, but resources like spell checker and tokenizer cannot be so easily adapted because the two languages use very different scripts.

Which resources can be adapted depends upon the kinds of similarities among those languages. Obviously, if two languages have similar morphology, a morphological analyzer developed for one can be adapted for the other. And if both the languages are free word order languages and have the same default word order (say, SOV) and also use postpositions to mark relations, then the parser built for one can possibly be adapted for the other. This is the case with languages like Hindi and Kannada, even though they belong to different linguistic families.

Thus, a strong claim can be made that measures of distances at a particular linguistic level can give us a reasonably good estimate of the cost of adapting a resource if that linguistic level is relevant for the concerned resource. A weak claim can also be made that the distance at a particular linguistic level can give us a rough estimate of the cost of adapting resources irrespective of the relevant linguistic level. The weak claim may be valid in many cases.

The cost of adaption also depends on the kind of resource, i.e., not just with respect to the relevant linguistic levels, but also on other characteristics of the resource. We will *not* be studying this second aspect of the cost of adaption. Therefore, once some estimate has been made of the cost of adaption based on the calculation of linguistic distance, the estimate may have to be revised to take care of other specific factors for the resource being adapted. Some of these factors might be completely non-linguistic, e.g. the practical constraints under which the resource developers will be working.

4. Objective and Subjective Estimation

As mentioned earlier, estimation of language distances can be either objective or subjective. The latter is needed when the former is not feasible or not sufficient. By objective we mean calculating the distance based on actual data (corpus) and using some computational technique without human involvement except in designing the method for computation. Subjective estimation, on the other hand, implies some human involvement in assigning numerical values based on

linguistic intuition, even during the process of estimation, not just while designing the method. The advantages and disadvantages of ‘objective’ and ‘subjective’ methods for our purpose are similar to those for other problem in computational linguistics, i.e., of using almost purely statistical method versus using almost purely linguistic methods. We will present examples of both objective and subjective estimation in the following sections.

4.1. Formulating the Problem

We would like to emphasize that the estimation of the cost of adaption, even when considering only language distance is not very amenable to mathematical formulation. In this section we try to formulate the problem in somewhat mathematical way just to ensure that the estimation is performed as objectively as possible, even if we use a ‘subjective’ method.

Let us say $P = \{p_1, p_2, \dots, p_m\}$ is the set of resource rich languages which can be potentially treated as pivot languages, i.e., their resources can be possibly adapted for some resource scarce languages $Q = \{q_1, q_2, \dots, q_n\}$. Further, $R = \{r_1, r_2, \dots, r_l\}$ is the set of resources we are interested in. A language can belong to both the sets, but not for the same resource, i.e., the intersection of the sets P and Q is empty if we are interested in only one resource, otherwise it may be non-empty because one language p_i may be lacking resource r_{k1} , while another language q_j may be lacking resource r_{k2} though not in r_{k1} .

If C_{ijk} is the cost of adapting the resource r_k of language p_i for language q_j , then the problem is to find out the pair p_i and q_j such that C_{ijk} is minimized:

$$\{p_i, q_j\} = \underset{i, j}{\operatorname{argmin}} C_{ijk} \quad (1)$$

As the above formulation implies, the problem involves the following steps:

1. Identify the languages which are potentially pivot languages (have some resources): P
2. Identify the languages which lack some resources: Q
3. Identify the resources which are relevant: R
4. Select some method for calculating the cost depending on the resources and the information available about the languages (corpora, grammar etc.).
5. Calculate the costs and select the pair which minimizes the cost.

Note that Q and R may be known *a priori* if we are only interested in building certain resources for certain languages. In such a case, the problem would be just to find p_i which will minimize the cost:

$$p_i = \underset{i}{\operatorname{argmin}} C_{ijk} \quad (2)$$

However, in the general case, this may not be true, e.g. when a large scale project is being launched to solve the problem of resource scarcity for as many languages as possible.

4.2. Objective Estimation

In this section we describe a method for objective estimation and present the results of some experiments using this method for many Indian languages or varieties. We will rely mainly on a corpus and the method itself would use linguistically informed measures of similarity.

4.2.1. Similarity Measures

We use two similarity measures for estimating the distance between languages. Both of these are designed to use some linguistic information at the level of writing system, phonology and lexicon (Singh and Surana, 2007). The basic idea is to first extract a list of highly frequent words from the unannotated corpus of each language. No other language resource is used. Since these words are highly frequent, they are likely to be from the core vocabulary of the language. This is in line with the insights gained from the techniques used in historical linguistics for comparing languages to find out whether they are related or not.

Then we use a method based on Computational Phonetic Model of Scripts (CPMS) to identify the cognates among these languages by calculating the surface similarity scores of pairs of words from the word lists extracted (Singh, 2006). By aligning word pairs using these scores and then using a threshold, we can identify the likely cognates, or more accurately, words of common origin since the method does not distinguish commonly inherited words from borrowed words.

The first measure of language distance (or of the cost of adaption of resources) is simply based on the idea that the more words of common origin the two languages have, the more likely they are to be similar. This measure is called the Cognate Coverage Distance (CCD). Cognate coverage distance gives us a measure of similarity of two languages, but it does so without taking into account the phonetic difference between two cognates. To include this factor, we use another measure called the Phonetic Distance of Cognates (PDC). Both of these measures have been previously used for calculation of distances (Singh and Surana, 2007) among languages and have shown very good correlation with purely linguistic subjective ‘estimates’ of linguistic distances among languages (in terms of their situation on the ‘tree’ of language families). Note that these previous experiments were performed using word lists extracted from corpus, not Swadesh-like word lists (Swadesh, 1952; Dyen et al., 1992) handcrafted by linguists.

The CPMS takes advantage of the fact that most of the major Indian languages use Brahmi origin scripts, which have a lot of similarities. It models all the similarities as well as other characteristics of the scripts such the close correspondence between letters and articulatory features. It uses a Stepped Distance Function (SDF) to calculate the distance between two letters in terms of phonetic (articulatory) and a few orthographic features. For calculating the distance between two words or strings, whether from the same language or from different languages, it uses an alignment algorithm which is an adapted version of the Dynamic Time Warping (DTW) algorithm (Myers and Rabiner, 1981). The CPMS based method for calculating surface similarity takes into account order sensitivity as well

as scaling (Heeringa et al., 2006; Ellison and Kirby, 2006). In general, the surface similarity of two strings can be defined as:

$$c_{lm} = f_p(w_l, w_m) \quad (3)$$

where f_p is the function which calculates surface similarity based cost between the word w_l of language l and the word w_m of language m .

The word pairs with the minimum cost are identified as cognates, provided that the cost is below a threshold. Once we have identified the possible cognates, we measure distance between two languages in two ways.

$$d_{pdc} = \frac{C'_{lm} + C'_{ml}}{2} \quad (4)$$

The first measure of language distance (or of the cost of adaption of resources) is simply based on the idea that the more words of common origin the two languages have, the more likely they are to be similar. This measure is called the Cognate Coverage Distance (CCD). Using CCD, the normalized distance between two languages can be defined as:

$$t'_{lm} = 1 - \frac{t_{lm}}{\max(t)} \quad (5)$$

where t_{lm} and t_{ml} are the number of cognates found when comparing from language l to m and from language m to l , respectively.

Since the CPMS based measure of surface lexical similarity is asymmetric, the average number of unidirectional cognates is taken as the distance:

$$d^{ccd} = \frac{t'_{lm} + t'_{ml}}{2} \quad (6)$$

Cognate coverage distance gives us a measure of similarity of two languages, but it does so without taking into account the phonetic difference between two cognates. To include this factor, we use another measure called the Phonetic Distance of Cognates (PDC).

This measure does not just calculate the coverage of cognates, but it also calculates the surface similarity based distance between each pair of cognates. The distance is the symmetric normalized sum of distances between cognate pairs. If n cognates are found between two languages, then:

$$C_{lm}^{pdc} = \sum_{i=0}^n c_{lm} \quad (7)$$

where n is the minimum of t_{lm} for all the language pairs compared.

The normalized distance is defined as:

$$C'_{lm} = \frac{C_{lm}^{pdc}}{\max(C^{pdc})} \quad (8)$$

Just as for CCD, a symmetric version of this cost can be calculated:

$$d_{pdc} = \frac{C'_{lm} + C'_{ml}}{2} \quad (9)$$

	AS	BP	GJ	KK	KN	ML	BM	MT	OR	RJ	PB	ST	TM
BN	0.00 0.21	0.74 0.34	0.59 0.50	0.68 0.66	0.61 0.41	0.77 0.50	0.83 0.39	0.49 0.39	0.01 0.05	0.65 0.42	0.57 0.48	0.80 0.67	0.91 0.73
HI	0.32 0.26	0.71 0.07	0.39 0.14	0.64 0.51	0.52 0.07	0.71 0.41	0.86 0.51	0.33 0.03	0.29 0.26	0.50 0.17	0.27 0.17	0.82 0.78	0.89 0.78
KN	0.63 0.39	0.82 0.17	0.69 0.26	0.62 0.55	- -	0.68 0.39	0.93 0.62	0.73 0.37	0.65 0.33	0.79 0.48	0.70 0.35	0.90 0.89	0.82 0.78
ML	0.76 0.50	0.90 0.42	0.80 0.62	0.77 0.75	0.68 0.39	- -	0.96 0.92	0.83 0.55	0.77 0.53	0.88 0.78	0.83 0.66	0.95 0.91	0.61 0.48
MR	0.55 0.41	0.80 0.26	0.52 0.14	0.65 0.64	0.57 0.12	0.78 0.55	0.92 0.71	0.63 0.37	0.54 0.35	0.70 0.44	0.55 0.28	0.89 0.91	0.91 0.78
TL	0.64 0.39	0.81 0.10	0.70 0.41	0.61 0.58	0.46 0.25	0.64 0.32	0.95 0.71	0.72 0.51	0.63 0.37	0.81 0.48	0.70 0.46	0.92 0.80	0.81 0.82
TM	0.90 0.82	0.96 0.73	0.94 0.91	0.79 0.91	0.82 0.78	0.61 0.48	0.98 0.89	0.91 0.85	0.93 0.76	0.92 0.75	0.87 0.75	0.95 0.94	- -

BN: Bengali, **HI:** Hindi, **KN:** Kannada, **ML:** Malayalam, **MR:** Marathi, **TL:** Telugu, **TM:** Tamil,
AS: Assamese, **BP:** Bhojpuri, **GJ:** Gujarati, **KK:** Konkani, **BM:** Bishnupriya Manipuri,
MT: Maithili, **OR:** Oriya, **RJ:** Rajasthani, **PB:** Punjabi, **ST:** Santali

Table 1: Distances between resource rich (rows) and resource poor (columns) varieties

4.3. Experiments

We conducted experiments on 17 varieties (languages or dialects). Out of these, four (Hindi, Bengali, Telugu and Marathi) are considered only as possible source (resource rich) languages. Three (Tamil, Malayalam and Kannada) are considered both as possible source languages as well as resource scarce languages. The reason for this is that these languages do not easily fit in either category: they have some resources, but they lack many others.

Ten varieties (Assamese, Bhojpuri, Gujarati, Konkani, Bishnupriya Manipuri, Maithili, Oriya, Rajasthani, Punjabi and Santali) are considered only as resource scarce languages as they hardly have any language resources. We had difficulty even in collecting unannotated corpus for many of the resource scarce languages like Maithili, Bhojpuri, Santali etc. Then there was the problem of encodings or notations in which the text is available.

For all the 17 varieties, we converted the text into UTF-8 format so that the CPMS based method could be applied and distances could be calculated using the measures (CCD and PDC) described earlier.

4.4. Results

The results shown in Table-1 and Figure-1 give the distances between the source languages and the resource poor languages based on the measures CCD and PDC.

As Figure-1 shows, most of the results are intuitively correct. For example, Hindi is found to be the best source language for Bhojpuri, Maithili and Rajasthani. All these three varieties are considered to be dialects of Hindi. Note that the purpose of these experiments is not just to find out the best source languages, but also to get a quantitative estimate of the cost of adaption of resources for the purpose of allocating finance, manpower, time etc. Another example is that Santali is shown to be closest to Bengali, although quantitatively it is not very close even to Bengali (as it belongs to a completely different language family than the other varieties).

Another observation that can be made from Figure-1 is that Tamil seems to be the ‘worst’ candidate for being the ‘source’ for resource adaption. This can be explained by the fact that Tamil is indeed (linguistically) more distant from the ‘target’ varieties being considered than the other ‘source’ varieties.

	Weight	KN	ML	TM	TL
DAG	3	2	2	2	3
DIM	3	2	0	2	2
FWO	4	3	3	2	3
DWS	3	1	4	3	1
CNT	2	1	1	1	1
CAT	2	1	1	1	1
CDT	2	1	1	1	0
CST	2	0	1	1	1
CGT	2	1	1	1	1
CIM	2	1	1	1	1
CLT	2	1	1	1	1
CVT	2	1	1	0	1
CAB	2	1	0	1	0

Table 2: Features (and their numerical values and weights) for subjective estimation for four close languages

5. Subjective Estimation

In subjective estimation, we may not use any measures that are applied on corpus. Instead, we do the following:

1. Select some linguistic features which are relevant to the resources being adapted.
2. For each language (or variety), subjectively assign some numerical value to the feature.
3. Calculate the distance between two languages using the numerical values of the feature. Different features

	KN	ML	TM
DAG	3	3	3
DIM	0	6	0
FWO	0	0	4
DWS	0	9	6
CNT	0	0	0
CAT	0	0	0
CDT	2	2	2
CST	2	0	0
CGT	0	0	0
CIM	0	0	0
CLT	0	0	0
CVT	0	0	2
CAB	2	0	2
Total	9	20	19

Table 3: Differences of weighted feature values for Kannada, Malayalam and Tamil from Telugu

can be given different weights depending upon their relevance.

The simplest way will be to select only boolean features or transform linguistic properties into boolean features. In such a case, assigning a numerical value (0 or 1) will be straightforward and so will be the calculation of distance. It is not required that all (or most of) the properties of languages be considered. We just have to pick up some representative and relevant (for resources) properties. There should be enough of them to give us a good estimate of the cost of adapting resources.

5.1. An Example

To demonstrate how subjective estimation can be performed, we will consider a hypothetical case where the resources we are interested in adapting are rule based morphological analyzer and parser. We assume that these are available for Telugu. The purpose is to build a parser for the other three South Indian (Dravidian) languages, namely Kannada, Tamil and Malayalam. We wish to calculate the relative costs of adapting the Telugu resources for these three languages.

In the first step, we define the following features, mostly based on Caldwell (Caldwell, 1913), principally because they represent relevant characteristics and can also be given numerical values:

1. **DAG**: Degree of agglutination
2. **DIM**: Degree of inflection with respect to number, person and gender
3. **DIT**: Degree of inflection with respect to tense, aspect and modality
4. **FWO**: The extent to which free word order applies
5. **DWS**: Degree of word segmentation
6. A set of features in which each (boolean) feature represents the presence of a particular case: nominative

(**CNT**), accusative (**CAT**), dative (**CDT**), sociative (**CST**), genitive (**CGT**), instrumental (**CIM**), locative (**CLT**), vocative (**CVT**) and ablative (**CAB**)

7. A set of (boolean) features for the types of pronouns: personal (**PPN**), adjectival (**PAV**) and relative (**PRT**)

Table-2 shows the features, the weights assigned to them, and their values for the four languages. Table-3 gives the weighted values of the features. As can be seen from this table, the relative costs of adapting resources of Telugu for Kannada, Malayalam and Tamil are 9, 20 and 19 respectively. These result seem to be intuitively correct.

6. Practical Problems in Estimation

Many practical problems can be encountered while estimating the cost of adaption, whether we use a subjective method or an objective method. For example, we had difficulty even in collecting unannotated corpus for many of the resource scarce languages like Maithili, Bhojpuri, Santali etc. Then there is the problem of encodings or notations in which the text is available. We also had to build encoding converters by manually preparing mappings for some of the languages. The text in Konkani was in an ASCII based notation, while that in Santali was in a different script called Ol Chiki. We could not use the Avadhi corpus (Bible) as the text was in PDF format and on saving it as text, it was saved in an encoding for which we couldn't prepare a converter. Even when enough text is available and it is possible to convert to a common notation or script, the characteristics of the corpora of different languages may be very different. For example, the text for Rajasthani was not just less in quantity, but was a mixture of several of its 'dialects', i.e., was not in standard Rajasthani. Note that Rajasthani is itself considered a 'dialect' of Hindi and is not a 'scheduled' or officially recognized language, even though it covers one of the largest areas in India.

In subjective estimation, the first problem is to select the features. The second problem is to select feature values. The third problem is to assign numerical values to feature values. The fourth problem is to assign weights to features. Then there is the problem of combining the differences in numerical feature values to arrive at a quantitative measure of the cost of adaption. All these problems require a subjective judgment, which can be made only by linguistically trained people who are familiar with the grammatical properties of the languages being considered.

Finally, there is the problem of combining the quantitative measures obtained from objective and subjective estimation and using it in making an estimate of the financial resources, the time and the manpower required.

7. Conclusion

In this paper we suggest that one of the ways to bridge the resource gaps among languages is to adapt the resources of a resource rich language for linguistically close resource poor languages. We have discussed how the cost of such adaption of resources can be estimated for large scale planning of resource building. Such estimation can be either objective or subjective. As an example of objective estimation, we presented the results of our experiments on some

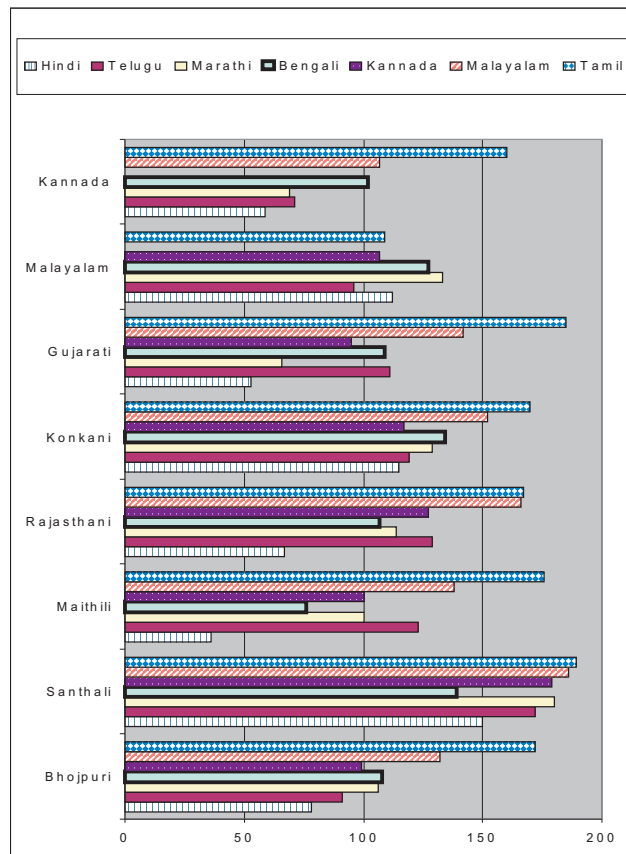


Figure 1: Selecting the best source language for some resource scarce languages. The length of the bar represents the distance and the cost of adaptation. The pattern or the shade in the bars represents a language.

languages and some major dialects which are considered by many to be languages. We also presented an example of subjective estimation.

8. References

- Robert Caldwell. 1913. *A Comparative Grammar of the Dravidian or South-Indian Family of Languages*. Kegan Paul, Trench, Trubner & Co. Ltd., London.
- I. Dyen, J.B. Kruskal, and P. Black. 1992. An indo-european classification: A lexicostatistical experiment. In *Transactions of the American Philosophical Society*, 82:1-132.
- T. Mark Ellison and Simon Kirby. 2006. Measuring language divergence by intra-lexical comparison. In *Proceedings of ACL*, Sydney, Australia. Association for Computational Linguistics.
- M. B. Emeneau. 1956. India as a linguistic area. In *Linguistics* 32:3-16.
- M. B. Emeneau. 1980. *Language and linguistic area. Essays by Murray B. Emeneau. Selected and introduced by Anwar S. Dil*. Stanford University Press.
- W. Heeringa, P. Kleiweg, C. Gooskens, and J. Nerbonne. 2006. Evaluation of String Distance Algorithms for Dialectology. In *Proc. of ACL Workshop on Linguistic Distances*.
- Grzegorz Kondrak. 2002. *Algorithms for language reconstruction*. Ph.D. thesis. Adviser-Graeme Hirst.
- April McMahon and Robert McMahon. 2005. *Language Classification by the Numbers*. Oxford University Press, Oxford.
- C. S. Myers and L. R. Rabiner. 1981. A comparative study of several dynamic time-warping algorithms for connected word recognition. In *The Bell System Technical Journal*, 60(7), pages 1389–1409.
- Luay Nakleh, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. pages 81–2:382–420.
- J. Nerbonne and W. Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of SIGPHON-97: 3rd Meeting of the ACL Special Interest Group in Computational Phonology*.
- J. Nerbonne. 2005. Review of ‘language classification by the numbers’ by april mcmahon and robert mcmahon.
- Anil Kumar Singh and Harshit Surana. 2007. Can corpus based measures be used for comparative study of languages? In *Proceedings of the ACL Workshop Computing and Historical Phonology*, Prague, Czech Republic.
- Anil Kumar Singh. 2006. A computational phonetic model for indian language scripts. In *Constraints on Spelling Changes: Fifth International Workshop on Writing Systems*, Nijmegen, The Netherlands.
- M. Swadesh. 1952. Lexico-dating of prehistoric ethnic contacts. In *Proceedings of the American philosophical society*, 96(4).