Cross-Entropy and Linguistic Typology

Patrick Juola
Department of Experimental Psychology
University of Oxford
Oxford, UK OX1 3UD
patrick.juola@psy.ox.ac.uk

Abstract

The idea of "familial relationships" among languages is well-established and accepted, although some controversies persist in a few specific instances. By painstakingly recording and identifying regularities and similarities and comparing these to the historical record, linguists have been able to produce a general "family tree" incorporating most natural languages.

We suggest here that much of these trees can be automatically determined by a complementary technique of distributional analysis. Recent work by (Farach et al., 1995) and (Juola, 1997) suggests that Kullback-Leibler divergence (or cross-entropy) can be meaningfully measured from small samples, in some cases as small as only 20 or so words. Using these techniques, we define and measure a distance function between translations of a small corpus (c. 70 words/sample) covering much of the accepted Indo-European family, and reconstruct a relationship tree by hierarchical cluster analysis. The resulting tree shows remarkable similarity to the accepted Indo-European family; this we read as evidence both for the immense power of this measurement technique and for the validity of this kind of mechanical similarity judgement in the identification of typological relationships. Furthermore, this technique is in theory sensitive to different sorts of relationships than more common word-list based methods and may help illuminate these from a different direction.

1 Introduction

Over the past century, a large amount of research effort has gone into the establishment of structures describing the typological and taxonomic relationships among languages past and present; the well-known "Romance language" group, consisting of all the languages in some sense "descended from" Latin is an example. In addition to their inherent interest, the results of these studies can be of use in telling us about the relationships, cultures, and environments of people and tribes long-distant from our present world.

Although these techniques are powerful, they are limited in their application in several ways. The traditional focus on word lists as the primary tool for language classification excludes syntax and morphology from consideration. By constructing these word lists out of only basic lexical items, the applicability is further limited. Although in theory these problems could be avoided by simply constructing different lists, there is still a problem with the volume of data to be processed — if the comparisons are performed at the level of "language," it is difficult if not impossible to discuss questions such as whether "legal English" shows more French influence than "standard English" or vice versa. However, the answers (were they available) to questions like this could be useful to, for example, sociolinguists in attempting to trace the relationships between and among subgroups within a culture.

The results presented in this paper suggest that distributional analyses can provide much of the same sort of relationships, but by a different route and therefore with different limitations and complementary to more standard techniques. This is developed further in a set of experiments which approximately reconstruct the accepted Indo-European family tree based on samples of running text of less than a page in length (and, in fact, typically under 70 words).

2 Taxonomy

Given the broad agreement found on the taxonomic relationships among languages [for example, see the introductory textbooks by (Gleason, 1955; Crystal, 1987; Finegan and Besnier, 1987), or the more authoritative (Bright, 1992; Asher and Simpson, 1994; Warnow, 1997)] the classifications and relationships of figure 1 can be described as uncontroversial. For example, the languages of Dutch and German are

Patrick Juola (1998) **Cross-Entropy and Linguistic Typology.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning,* ACL, pp 141-149.

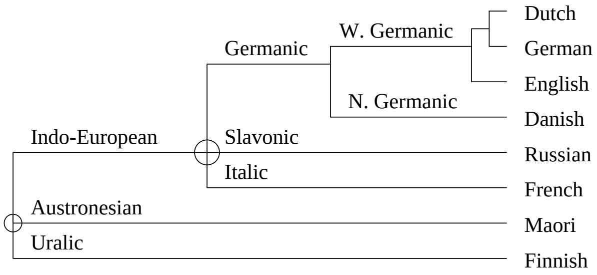

Figure 1: Genetic taxonomy of various languages

rather self-evidently similar; they are also closely linked in terms of history, culture, and linguistic borrowing; this similarity is one of the sources of evidence for such linkages. Meanwhile, there's little or no evidence that the Germans and the Maori were ever in significant day-to-day contact, a judgement borne out by apparent dissimilarity. The most controversial point of the diagram, as a matter of fact, may be its tree-like structure, as will be discussed later.

The usual method for generating such trees (or other representational structures) is to painstakingly compare representative samples of language, usually lists of lexical items, and identify similar or isomorphic changes from among the lists (taking into account historical and archeological evidence as appropriate). (Swadesh, 1955), for example, has identified a hundred basic concepts that are, in theory, part of the basic vocabulary of a language and thus resistant to borrowing and replacement and subject only to the slow "evolutionary" pressures of linguistic change. By comparing the presentation of these concepts as lexical items and measuring the degree of change between two languages' presentations, one can determine the amount by which two languages have "drifted."

In summary of the results of these and similar studies, (Finegan and Besnier, 1987) identify no less than eleven subgroups within the Indo-European family. In addition to the well-known groups like Germanic, Italic, and "Slavonic" (described here), they list Albanian, Anatolian, Armenian, Baltic, Celtic, Greek, Indo-Iranian, and Tocharian. (Crystal, 1987) groups Baltic and Slavic but otherwise agrees with Finegan and Besnier, as does (Gleason, 1955). This shows both the power of this technique as well as the degree to which it requires subjective evaluation; the overall relationships are generally agreed upon, but "the devil is in the details" and opinions about exactly which changes are similar remain to a certain extent educated guesses.

Other minor problems with this technique exist; for example, Swadesh's vocabulary list is completely insensitive to other aspects of language such as morphology, syntax, and so forth. Because of its focus on specific, basic words, it can be trapped (or tricked) by lexical drift (for example, "meat" is no longer the English word for "any foodstuff") or lexical holes where a clear cognate is not necessarily the most common or most frequent lexeme ((Forster et al., in press) has found that some of his Alpine languages have no lexeme for "to sit," for example.) Similar problems exist with regard to lexical borrowing; resistant to borrowing does not equate to proof against borrowing. Finally, this focus on these very basic terms and the evaluation of language as a whole may, to a certain extent, preclude the analysis of the paths of borrowing and the degree to which linguistic change is confined to or driven by particular fields, social strata, and so forth. By confining ourselves to pre-set lists of specific concepts, one runs the risk of picking the wrong concepts, especially for specific sub-fields (which can be as finely subdivided as one likes; is this paper an example of "science," of "computer science," of "computational linguistics," or of "information-theoretic approaches to corpus-based computational linguistics"?) As a simple example, the phrase for "TCP/IP protocol" in most languages of the world is recognizably a borrowing from English, while much of the jargon in the martial arts community shows a strong Japanese influence, even when the martial art itself derives from other countries or cultures.

This suggests that there is a place for other measures, both of language-in-use and of smaller samples, as a supplement to traditional typological and taxonomic measures. The claim made here is that cross-entropy (or Kullback-Leibler divergence) can be the basis for such a measurement.

3 Entropy Estimation

3.1 Background

English, as is well-known, is very predictable. Fluent English readers can confirm this for themselves by guessing which letter comes next in a word beginning *psyc-*. Experiments by (Shannon, 1951) indicate that most readers can guess more than half of the letters in running text based on their expert knowledge of the lexicon, structure, and semantics of English.

This notion of predictability, as well as the associated concepts of complexity, compressiveness, and randomness, can be mathematically modelled using information entropy. As developed by (Shannon, 1948), the entropy of a (stationary, ergodic) message source is the amount of information, typically measured in bits (yes/no questions), required to describe the successive messages emitted by that source to a recipient. As the set of possible messages becomes larger, or the distribution of messages becomes less predictable, the entropy of the source increases correspondingly, in accordance with Shannon's equation:

$$H(P) = -\sum_{i=1}^{N} p_i \cdot \log_2 p_i \tag{1}$$

where P is (the probability distribution of) a source capable of sending any of the messages $1, 2, \ldots, N$, each with some probability p_i. (For continuous distributions, simply replace the summation with the appropriate integral.)

An important aspect of this brief description has significant typological and taxonomic implications. Against what is the predictability of the distribution measured? The second term in the above equation is a measure of the efficiency of the representation of message i (obviously, more frequent messages should be made shorter for maximal efficiency, an observation often attributed to Zipf), based on our estimate of the frequency with which i is transmitted. Therefore, we can generalize equation 1 to

$$\hat{H}(P, Q) = -\sum_{i=1}^{N} p_i \cdot \log_2 q_i \tag{2}$$

where Q is a different distribution representing our best estimate of the true distribution P. This value (called the cross-entropy) achieves a minimum when $P = Q$, and $\hat{H}(P, P) = H(P)$. The difference between $\hat{H}$ and H , the so-called Kullback-Leibler divergence, can be taken as a measurement of the degree of similarity between P and Q.[1] For further elaboration on this point, the reader is referred to the excellent treatment in (Bishop, 1995).

This technique lends itself to a measurement of similarity between two different sources, by estimating the distributional parameters and calculating their cross-entropy.

3.2 Method

Obviously, much research has been done in the proper development of distributional models of English (or other languages) and in the efficient estimation of the probability distribution; (Brown et al., 1992) calculate the entropy of a statistical model of English that was produced by training a computer on literally billions of observations comprising a huge corpus of written English. (Wyner, in press) has suggested that one can determine the entropy to nearly as good accuracy based on much smaller sample sizes, but it remains an open research question how much text is actually needed. At billions of observations per test, it is obviously impractical to determine document-level properties (such as, for instance, authorship, register, difficulty of reading, or even the language in which a novel document is written), but if the tests can be made sufficiently sensitive to work with small texts, tests like this may be practical.

(Farach et al., 1995; Wyner, in press) describe a novel algorithm for entropy estimation for which they claim very fast convergence time; using no more than about five pages of text, they can achieve nearly the same accuracy as (Brown et al., 1992). The heart of this technique is a measurement of "match length within a database." Wyner defines the match length $L_n(x)$ of a sequence $(x_1, x_2, \ldots, x_n, x_{n+1}, \ldots)$ as the length of the the longest prefix of the sequence $(x_{n+1}, \ldots)$ that matches a contiguous substring of $(x_1, x_2, \ldots, x_n)$, and proves that this converges in the limit to the value $\frac{\log n}{H}$ as n increases.

A simple example should make this more clear : we consider for a moment the phrase

HAMLET : TO BE OR NOT TO BE THAT IS THE QUESTION

and fix n at 21. Thus, the "database" is the characters "HAMLET : TO BE OR NOT" (length 21) and the string " TO BE THAT IS THE QUESTION" is the remaining data; the prefix " TO BE " exactly matches the contiguous substring beginning of the eighth character and itself runs for seven characters, but the prefix " TO BE T" does *not* match a continuous substring of the database, and hence the match length L_{21} is seven.

Using this technique, one can estimate the entropy of a sequence by sliding a block of n observations along the sequence and calculating the the mean match length $\hat{L}$ (averaged over each step) and thus the estimated entropy $\hat{H}$. So one calculates L_{21} above, then calculates L_{21} for the string "AMLET : TO BE OR NOT TO BE THAT IS THE QUESTION ", then for "MLET : TO BE OR NOT TO BE THAT IS THE QUESTION W", and so on.

The application of this to measurement of cross-entropy is relatively straightforward. A "database" of n observations is compiled for each language of interest and each successive symbol of the message stream of interest is used as the starting point for

[1] N.b. this is not a "distance metric" in the formal sense of the word (it's not symmetric, for one thing), but can be thought of as a distance for these purposes.

the maximal prefix to be found within the database. Although this loses some of the time-varying properties of an entropy estimator (in particular, the database is fixed and will not shift to capture long-term regularities in an input stream), this should preserve the fundamental relationship that a closer fit (smaller cross-entropy) results in a longer mean match length. This permits us to measure cross-entropy with approximately the same convergence properties as the entropy estimation itself.

The primary claim made in this paper is that the similarity measured by cross-entropy will have some of the same properties for typological and taxonomic research as those of more conventional word-lists, but that cross-entropy is complementary in several ways. It is easier and more accurate to measure cross-entropy in this way, is sensitive to the sublanguage of the samples used (and hence can be used for smaller-scale experiments), and is sensitive to aspects of language, such as syntax, lexical choice, and style, that are not commonly found in word lists. For example, languages with similar lexical items but different structures (perhaps verb-medial instead of verb-final) will find fewer multi-word matches between the databases, and thus will produce a greater measured distance, indicative not of the lexical distance but of the *syntactic.*

3.3 Corpora

Several experiments have been performed to test this hypothesis. The first, detailed in (Juola, 1997) simply approaches this as a language-identification problem. Given a set of linguistic samples (in this case, Danish, Dutch, English, French, German, and Spanish, plus, as distractors, Finnish, Finnish, and Maori) in which of the sampled languages was a novel text written? Using samples of 100, 250, and 500 characters, 472 documents, ranging in size from <500 to several million characters. The remarkable accuracy possible, even with very small samples, is shown by the fact that, for instance, at the 250 character level, only one document was miscategorized (German misclassified as Dutch), even when texts to be identified were from completely separate registers.

The second experiment involved the languages described in figure 1. Samples of 1000 characters from the beginning of the book of Genesis were taken from each of the languages (the Russian sample being automatically transliterated into a Latin-character "equivalent") and cross-entropy between each pair (e.g. how close German is to the Dutch database) was measured. These pairs were averaged (n.b. the cross-entropy between Dutch and German is not necessarily the same as the cross-entropy between German and Dutch) to produce a symmetric "distance" matrix, and agglomerative cluster analysis was performed to produce set of binary "tree" relationships. This analysis consisted of simply taking all pairwise distances, and making a "cluster" of the two clusters with the smallest minimum (mean, or maximum) distance and continuing until the entire set was combined into a single cluster. (Obviously, these might produce three slightly different trees; results reported here are from the minimum tree throughout.)

Please read the following *aloud*:
I hereby undertake not to remove from the Library, or to mark, deface, or injure in any way, any volume, document, or other object belonging to it or in its custody; not to bring into the Library or kindle therein any fire or flame, and not to smoke in the Library; and I promise to obey all the rules of the Library.

Figure 2: Bodleian declaration in English

The third experiment was similar to but broader than the second. For the past several decades, an informal project of the Bodleian Library, Oxford, has been the gathering of translations of the traditional declaration to be taken by all new members of the University (and others) before access can be granted to the books. As a convenience to the international community of scholars, the librarians have attempted to gather translations of this declaration in as many languages as possible so that scholars can be made aware of what they are promising; as a goal, they have set for themselves the task of acquiring the declaration both in every language spoken in Europe (including some nearly "dead" languages such as Cornish and Breton) as well as in at least one official language for every country in the world (or at least every country represented at the United Nations). The definitive version of the declaration is the one in English, reproduced here as figure 2; also reproduced is the translation into Basque.

From this collection were taken samples of fifty-three languages, mostly spoken in Europe or derived from European languages (n.b. not necessarily of the Indo-European family, e.g. Basque and Maltese) and written primarily in the standard Latin script. These samples typically range between 300-400 characters each. As before, cross-entropy measurements were taken (and symmetrized) between every pair and used as the basis for an agglomerative cluster analysis.

We expect, of course, in the second and third ex-

Agintzerakoan, adierazpen hau irakur ezazu mesedez, *ahots goraz*:
Honen bidez Liburutegiari dagozkion liburuki, eskribu, edo beste inolako gauzarik ez eraman, ez markatu, ez hondatu, edo beste edozein moduzko kalte ez dudanik egingo hitz ematen dut; Liburutegi barnean ez erre, ez piztu, ezta beste inolako sua sartu, eta Liburutegiko araudi guziak obedituko ditudala hitz ematen dut.

Figure 3: Bodleian declaration in Basque

Afrikaans, Albanian, Basque, Breton, Catalan, Cornish, Croatian, Czech #1, Czech #2, Danish, Dutch, English (Middle), English (Modern), English (Old), Esperanto, Estonian, Faeroese, Finnish, French, Frisian, Galacian, German, Hungarian, Icelandic, Irish (Gaelic), Italian, Ladin (Dolomitic), Ladin (Friulan), Ladin (Romontsch), Lappish, Latvian, Lithuanian, Macedonian, Maltese, Manx, Norwegian, Occitan, Polish, Portuguese, Provençal, Roumanian, Scots English, Scottish (Gaelic), Serbo-Croat, Slovak, Slovenian, Sorbian, Spanish, Urban Suebian, Swedish, Welsh

Figure 4: List of languages studied

periments that known linguistic groupings (such as Romance, Germanic, Slavic, and so forth) would appear as clusters within the final tree.

4 Results

As alluded to earlier, the results from the first experiment indicate that as few as 100 characters can be sufficient to identify the language in which a document is written; (Juola, 1997) contains more details.

Within the limitations of binary branching imposed by the cluster analysis algorithm, the family tree of figure 1 was reproduced perfectly in the second experiment; the circled nodes are, of course, ternary in this figure but binary in the recovered tree. The experimental results show that, instead of ternary branching, Maori is considered to be more distant from the Indo-European cluster than is Finnish and that (transliterated) Russian is more distinct from the Germanic cluster than is French; these findings, although not necessarily convincing from the standpoint of statistical significance, are certainly intuitively plausible given the geographic closeness and ease of communication and therefore linguistic borrowing. On the other hand, (Warnow, 1997) claims a greater degree of similarity between Slavic and Germanic languages than between Slavic and Romance; this discrepancy may simply reflect the accuracy limits of the corpus sizes used or may be evidence of a greater degree of *cultural* influence on Germany from the West than from the East which is not reflected in the basic vocabulary.

The results of the third experiment are less perfect, but in many regards more interesting. In general, the best results were obtained at what might be called "mid-level" regularities. (For simplicity, we concentrate here on the results of the minimal distance cluster analysis.) For example, all the languages of the Iberian peninsula (Galacian, Portuguese, Occitan, Catalan, and Spanish) were grouped into one tree, which was attached to *two* of the three Ladin samples (Friulan and Romontsch) but not to Dolomitic Ladin, a result compatible with the findings of (Forster et al., in press) that the level of linguistic diversity within the "Alpine Romance" languages is as great as the difference between, e.g. French and Italian. This cluster itself can be extended to incorporate all the Italic/Romance languages *except* Latin itself; again, this is compatible with the findings of (Forster et al., in press), and plausible in itself if one assumes that it's more useful for a speaker of modern Ladin to be able to understand modern Italian than classical Latin.

Similarly, (some of) the North Germanic languages (Danish, Norwegian, and Swedish) were clustered, as were the South Germanic languages Afrikaans, Dutch, German, Luxemburgish, and Frisian — but these two groups were themselves separated, with Danish *et al.* being measured as being closer to the Romance cluster than to the South Germanic. Similarly, the different varieties of English were widely separated, with Modern English, (Modern) Scots English, and Middle English being an identifiable cluster, but with Old English being grouped with Icelandic and Faeroese in a cluster distant from anything else.

The complete tree which the computer generated is attached on the following page. Each leaf is labeled with the appropriate language and with the subfamily of Indo-European from which it derives. Non-Indo-European languages, such as Basque or Finnish, are labelled with their families (in parentheses). All labels are to be regarded as largely consensual and representing common opinions, rather than as necessarily authoritative statements; in some cases, even the existence of languages (e.g. Croatian vs. Serbo-Croatian) can be divisive, as much for political and nationalistic as for scientific reasons.

5 Discussion

The results presented above, while preliminary (as a result of the small number of languages on the

```
        +- Basque (isolate)
      +-+
      | |     +- Cornish:Celtic
      | |   +-+
      | |   | +- Manx:Celtic
      | | +-+
      | | | |         +- Estonian (Finno-Ugric)
      | | | |       +-+
      | | | |       | +- Finnish (Finno-Ugric)
      | | | |     +-+
      | | | |     | |   +- Breton:Celtic
      | | | |     | | +-+
      | | | |     | | | |       +- Czech#1:Slavic
      | | | |     | | | |     +-+
      | | | |     | | | |     | | +- Czech#2:Slavic
      | | | |     | | | |     | +-+
      | | | |     | | | |     |   +- Slovak:Slavic
      | | | |     | | | |   +-+
      | | | |     | | | |   | +- Sorbian:Slavic
      | | | |     | | | | +-+
      | | | |     | | | | | |         +- Afrikaans:S. Germanic creole
      | | | |     | | | | | |       +-+
      | | | |     | | | | | |       | +- Dutch:S. Germanic
      | | | |     | | | | | |     +-+
      | | | |     | | | | | |     | | +- German:S. Germanic
      | | | |     | | | | | |     | +-+
      | | | |     | | | | | |     |   +- Luxemburgish:S. Germanic
      | | | |     | | | | | |   +-+
      | | | |     | | | | | |   | +- Frisian:W. Germanic
      | | | |     | | | | | | +-+
      | | | |     | | | | | | | |       +- Albanian:Albanian
      | | | |     | | | | | | | |     +-+
      | | | |     | | | | | | | |     | +- Maltese (Semitic)
      | | | |     | | | | | | | |   +-+
      | | | |     | | | | | | | |   | | +- Roumanian:Italic
      | | | |     | | | | | | | |   | +-+
      | | | |     | | | | | | | |   |   |   +- French:Italic
      | | | |     | | | | | | | |   |   | +-+
      | | | |     | | | | | | | |   |   | | |   +- Italian:Italic
      | | | |     | | | | | | | |   |   | | | +-+
      | | | |     | | | | | | | |   |   | | | | |     +- Galician:Italic
      | | | |     | | | | | | | |   |   | | | | |   +-+
      | | | |     | | | | | | | |   |   | | | | |   | | +- Portuguese:Italic
      | | | |     | | | | | | | |   |   | | | | |   | +-+
      | | | |     | | | | | | | |   |   | | | | |   |   | +- Occitan:Italic
      | | | |     | | | | | | | |   |   | | | | |   |   +-+
      | | | |     | | | | | | | |   |   | | | | |   |     | +- Catalan:Italic
      | | | |     | | | | | | | |   |   | | | | |   |     +-+
      | | | |     | | | | | | | |   |   | | | | |   |       +- Spanish:Italic
      | | | |     | | | | | | | |   |   | | | | | +-+
      | | | |     | | | | | | | |   |   | | | | | | | +- Ladin(Friulan):Italic
      | | | |     | | | | | | | |   |   | | | | | | +-+
      | | | |     | | | | | | | |   |   | | | | | |   +- Ladin(Romontsch):Italic
      | | | |     | | | | | | | |   |   | | | | +-+
      | | | |     | | | | | | | |   |   | | | |   +- Provencal:Italic
      | | | |     | | | | | | | |   |   | | +-+
      | | | |     | | | | | | | |   |   | |   +- Ladin(Dolomitic):Italic
      | | | |     | | | | | | | |   |   +-+
      | | | |     | | | | | | | |   |     | +- Esperanto:Italic artificial
      | | | |     | | | | | | | |   |     +-+
      | | | |     | | | | | | | |   |       | +- Lithuanian:Baltic
      | | | |     | | | | | | | |   |       +-+
      | | | |     | | | | | | | |   |         |   +- Croatian:Slavic
      | | | |     | | | | | | | |   |         | +-+
      | | | |     | | | | | | | |   |         | | +- Serbo-Croat:Slavic
      | | | |     | | | | | | | |   |         +-+
      | | | |     | | | | | | | |   |           +- Macedonian:Slavic
      | | | |     | | | | | | | | +-+
      | | | |     | | | | | | | | | | +- Danish:N. Germanic
      | | | |     | | | | | | | | | +-+
      | | | |     | | | | | | | | |   | +- Norwegian:N. Germanic
      | | | |     | | | | | | | | |   +-+
      | | | |     | | | | | | | | |     +- Swedish:N. Germanic
      | | | |     | | | | | | | +-+
      | | | |     | | | | | | |   +- Slovene:Slavic
      | | | |     | | | | | +-+
      | | | |     | | | | |   +- Latin:Italic
      | | | |     | | | +-+
      | | | |     | | |   +- Latvian:Baltic
      | | | |     | +-+
      | | | |     |   |   +- English(Modern):W. Germanic
      | | | |     |   | +-+
      | | | |     |   | | +- ScotsEnglish:W. Germanic
      | | | |     |   +-+
      | | | |     |     +- English(Middle):W. Germanic
      | | | |   +-+
      | | | |   | +- Polish:Slavic
      | | | | +-+
      | | | | | +- Hungarian (Finno-Ugric)
      | | | +-+
      | | |   | +- IrishGaelic:Celtic
      | | |   +-+
      | | |     +- ScottishGaelic:Celtic
      | +-+
      |   +- Lappish (Finno-Ugric)
    +-+
    | +- UrbanSuebian:Germanic dialect
  +-+
  | +- Welsh:Celtic
+-+
  | +- English(Old):W. Germanic
  +-+
    | +- Faeroese:N. Germanic
    +-+
      +- Icelandic:N. Germanic
```

one hand, and the small samples on the other), are promising; mid-range similarities, which might be independently expected to be the most stable, are indeed picked up with remarkable accuracy. Very subtle and distant relations are more likely to be masked by simple noise or random chance (cf. (Ringe, 1992)), while closely similar languages may be so similar that lexical choice and style, in some cases of a single word (do I describe something as "big" or "large"?), may be enough to alter the very closely-knit relationships. (For example, the two Czech samples are *not* sisters, but aunt/niece, as the Slovak sample intervenes — however, the Czech/Slovak samples themselves form a cluster.) Both of these effects can be expected to be reduced as the sample sizes increase; the primary finding that a few hundred characters of language *in use* can discover many of the relationships captured by more traditional methods in a numerical and objective way, avoiding the difficulties of interpreting whether two differences are "similar."

One major point of controversy will undoubtedly be the use of a tree structure for describing these relationships. There are, of course, two major models for describing linguistic families, the "tree" model and the "wave" model, and although (Warnow, 1997) may claim that the tree model is universally accepted except in cases of extremely closely related languages, this statement seems more firm than absolutely justified. However, the tree structure presented here is more an artifact of the cluster analysis technique used (and certainly the forced binary branching is artifactual) than a property of the entropy measurement technique.

One significant problem which has not been addressed entirely is the question of alphabet effects. First, the very idea of evaluating linguistic similarity by examination of letters, instead of sounds, will strike a traditional comparativist as almost nonsensical. Letter comparisons will only work to the extent that correspondence in written form reflects regularities in linguistic forms. Fortunately, the letter/sound correspondence for most languages, and particularly for most alphabetic languages, is significantly better than random, if not quite perfect. Comparisons between languages using different alphabets (for example between (Cyrillic) Russian and (Latin) English) produce uniformly and unsurprisingly huge differences.

The work presented here restricts itself almost entirely to languages written in the conventional Latin alphabet (with occasional diacritical mark or unusual character such as the Icelandic eth). However, even within this subset, focusing on written characters, as opposed to sounds, can change the similarity metrics. In some cases, the letter/letter similarity can actually be better than the sound/sound similarity, for example in cases where accents have drifted while the written form has been stabilized (e.g. consider the English, American, and Australian pronunciations of the word "grass"), or in cases where particular words have been borrowed but have had their pronunciation regularized to a local standard. In other cases, however, the same sound may be represented by different characters (the German 'W' vs the English 'V', or the Old English thorn, transcribed in modern English as the digraph 'th'). A particularly problematic area can be in the representation of diacritical marks – intuitively, one would expect that the letters ö and o would be somehow more similar than the letters e and o (or than t and o), particularly when one is considering words that may have been explicitly borrowed and lost their diacritics in the process).

In either of these instances, the borrowing itself can be read as evidence of cultural contact, possibly in connection with geographic proximity. In this case, the difference in apparent similarity between word-list methods (which presumably measure more of the historical relationships of descent and derivation) and the proposed method (which incorporates measurings of borrowing, and so forth) can be used as a complementary technique to measure such things as the rate, source, and paths of borrowing. In particular, measuring letter/letter as well as sound/sound differences might be a useful additional source of information for comparativists.

The possibility of two letters (or sounds) being "more similar" should also not be discounted (as has been done in this work). It was suggested above that ö and o are a "similar" letter pair; one would also expect that, for instance, /f/ and /v/ are "similar", especially in words borrowed into a language that doesn't have unvoiced consonants – while /f/ and /g/ would be almost universally distinct. By treating individual words/sounds as distinct, orthogonal, and *unanalyzed* symbols, the current technique may lose this sort of information in its measurements.

On the other hand, this sort of measurement explicitly allows document and subject level distinctions to be observed and validated. It is a commonplace observation, for example, that there is a greater preponderance of Latin- and Greek- based words in (English) scientific discourse than in general conversation; this is not especially based on any particular difference in the choice of lexical items, but more generally on the subject of discourse and the fact that the lexical items available for scientific discussions tend to be Latinate as opposed to Anglo-Saxon. (In other words, you can choose any word you like from the standard list – all of which are Latin-derived.) Thus, word-list based methods are unable to validate this distinction, and some other method such as comparative etymology might be required. Again, the proposed method can be used to determine complementary information to that gained via traditional techniques; the observation of the Latineque words in scientific, but not conversational, English will quite reasonably support the inference that scientists (or the group that gave rise to modern scientists) are more likely to have been exposed extensively to Latin than the general public, and thus that knowledge of Latin was characteristic of that particular segment of society.

6 Future Work and Conclusions

One obvious aspect of the Bodleian corpus is that, by construction, all items are translations of each other (or more accurately of the English). The acquisition of translated corpora in a sufficiently varied set of languages can be problematic; it would obviously be useful to test to what extent cross-entropy can be used as a taxonomic relationship on related corpora that are not necessarily translations of each other. Similarly, much further work is required to determine the best method of analysis, whether by cluster analysis or other techniques, and what degree of accuracy can be expected with various corpus sizes, registers, &c. On the other hand, if it's hard to acquire small translated corpora, it's even harder to acquire large ones, and the sensitivity of Wyner's entropy estimation technique is an undoubted advantage.

Further research will also be required to determine when to *stop* proclaiming relationships. As has been argued by (Ringe, 1992), the mere fact that two structures are similar does not imply that they are related; similarity may arise through mere chance. Given a reasonable model of language, it should be possible to determine what level of cross-entropy chance should predict, and thus when to stop agglutinating languages into proto-World and beyond, or determining whether a particular piano sonata should be classified as closer to Indo-European or Sino-Tibetian.

Going further afield, once the possibility of producing document, instead of language, taxonomies is accepted, it is possible to discuss meaningfully and to consider concepts such as the rate of change of a language (did English change more between 1600-1650 than between 1900-1950?) or the varying degrees of taxonomic relationships between various stylistic or subject classes. More generally, this cross-entropic method provides a way of combining information about relationships from a variety of sources, including lexical availability, lexical choice, pronuncations, syntax, and so forth.

Ultimately, cross-entropy will probably not replace the word-list differentiation method of determining historic and familial relationships between languages, but can provide a valuable supplement to more traditional methods, as well as being able to address questions that are currently unanswerable by standard methods. Cross-entropy appears to be a meaningful and easy to measure method of determining "linguistic distance" that is more sensitive to variances in lexical choice, word usage, style, and syntax than conventional methods. Furthermore, this allows scientists to study taxonomic relationships among much smaller samples of language than were previously possible and to provide some sort of numerical validation (to be confirmed or rejected). Although much further work is necessary to determine the exact limitations of this sort of similarity measurements, preliminary results indicate that the accepted taxonomy is nearly reconstructable from remarkably little corpora, which shows at least in principle the power of this technique.

7 Acknowledgements

This work was funded primarily by ESRC grant 70. The author would also like to thank Dr. John G. Pusey, Admissions Officer at the Bodleian Library, for making the Bodleian corpus available; should anyone wish to assist in this project (a wish with which the author heartily concurs), please contact Dr. Pusey at *admissions@bodley.ox.ac.uk* or at the Bodleian Library, Broad Street, Oxford, UK. The author would also like to acknowledge the valuable contributions of Jodi Affuso in transcribing the corpus onto disk, of Alex Popiel for his programming expertise, and of Todd Bailey and Anna Morpurgo-Davies for critical reading and discussion of the manuscript.

References

Ronald Eaton Asher and J. M. Y. Simpson, editors. 1994. *The Encyclopedia of Language and Linguistics.* Pergamon, Oxford.

Christopher M. Bishop. 1995. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

William Bright, editor. 1992. *International Encyclopedia of Linguistics.* Oxford University Press, Oxford.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jennifer C. Lai, and Robert L. Mercer. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1).

David Crystal. 1987. *The Cambridge Encyclopedia of Language*. Cambridge University Press, Cambridge, UK.

Martin Farach, Michiel Noordewier, Serap Savari, Lary Shepp, Abraham Wyner, and Jacob Ziv. 1995. On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In *Proceedings of the 6th Annual Symposium on Discrete Algorithms (SODA95)*. ACM Press.

Edward Finegan and Niko Besnier. 1987. *Language, Its Structure and Use*. Harcourt Brace Jovanovich, San Diego.

Peter Forster, Alfred Toth, and Hans-Juergen Bandelt. in press. Phylogenetic network analysis of word lists. *Journal of Quantitative Linguistics*.

H. A. Gleason. 1955. *Introduction to Descriptive Linguistics*. Holt, Rinehart and Winston, New York.

Patrick Juola. 1997. What can we do with small corpora? Document categorization via cross-entropy. In *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*, Edinburgh, UK. Department of Artificial Intelligence, University of Edinburgh.

Donald A. Ringe. 1992. *On calculating the factor of chance in language comparison*, volume 82 of *Transactions of the American Philosophical Society*. American Philosophical Society.

Claude Elmwood Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Claude Elmwood Shannon. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatic dating. *International Journal of American Linguistics*, 21:121–37.

Tandy Warnow. 1997. Mathematical approaches to comparative linguistics. *Proceedings of the National Academy of Sciences of the USA*, 94:6585–90.

Abraham J. Wyner. in press. Entropy estimation and patterns.