Experiments in Unsupervised Learning of Natural Language

David Powers
Dept of Computer Science
Flinders University of South Australia
powers@cs.flinders.edu.au

Summary

We discard all theories of syntax and grammar whilst retaining the linguistic and scientific methodologies which have lead to their postulation. We add to our recipe ideas from psycholinguistics and information theory, the availability of many corpora of raw text, and the challenge of dealing with spoken language, and seek to learn language the way linguists, scientists and/or babies do.

The work reported in this paper concentrates on structure below the word level.

Keywords

Unsupervised Machine Learning/Natural Language Learning/Poverty of the Stimulus/Self-Organization

Introduction

Linguistics has invented and discarded many theories of language, and there are currently many competitors to the basic idea of phrase structure grammars as capturing the syntactic structure of language. Computational Linguistics has proven to be a testing ground for theories and grammars, and is similarly diverse. Moreover recently we have learnt that the similar principles and techniques may be applied at different linguistic levels, including morphology, nd phonology and orthography.

Whilst many of these approaches have proven useful in diverse ways, none have any substantive claim to reflect the way human language is actually represented, structured and processed in the human brain, although there is relevant work in Neurolinguistics, Psycholinguistics and Cognitive Linguistics.

We explicitly reject the Generative framework, the Poverty of the Stimulus argument, and the nativist Principles and Parameters theory. As argued extensively elsewhere [Powe89], we adopt a form of Piagetian constructivist framework in the context of a Cognitive Linguistic perspective on language using a Tagmemic approach to analysis [Pike77] which is inspired by Phonemics [Pike47].

Avoiding commitment to a grammatical theory, we have also discarded the role of word and sentence as givens in our grammar, and question even the supposed recursive basis of language. However, this apparent radicalism has grown out of our preliminary experiments [Powe89] and reflects the lack of definition of what we mean by these terms, and the lack of precise correspondence between the linguistic levels associated with the terms in different languages.

We commence by reviewing these original unsupervised learning experiments of [Powe89] and the observations which emerged from them. We then turn to the more radical experiments of [Powe91] and our subsequent generalization of these into a hierarchical learning paradigm.

In the body of the paper, we present ongoing experiments which aim to evaluate different choices of algorithm and metric, and extend the paradigm in directions which are designed to overcome the present limitations. These are presented in the rather unglamorous context of the first stage of classification or as an intuitive multidimensionally scaled map, and later in dendrograms under additional assumptions of hierarchical construction. This is part of a large evaluation of techniques preliminary to undertaking further experiments at the syntactic and morphological levels. Experiments have also been performed to do phonemic and syllabic classification based on phonetic data, but for simplicity we use orthographic representations only in this paper. The technique can also demonstrate significant utility here.

Finally, we seek to draw conclusions about the utility of different metrics, transformations and clustering assumptions, and in passing we show how this applies to syntax and relate the kind of grammars learned to traditional phrase-structure and feature-grammar approaches.

Learning Paradigms

Machine Learning of Natural Language prior to the 90s has been reviewed in [Powe89,91] and [Lang91]. Up to this point, it can be characterized as being highly supervised, with typically a sentence and its target semantic representation being provided, and/or explicit interaction on the acceptability of

proposed rules - some of the experiments in [Powe89] had this characters too. Unfortunately this highly interactive approach to learning is not that useful, as it invariably assumes that we know the desired structures and/or rules. But if we knew them, NLP would be rather different...

In the 1990s, connectionist approaches to NL have become important, as evidenced by several workshops. More recently, statistical corpus linguistic has moved from seeking merely to improve the performance of a (speech) system to providing the heart of the system. Interestingly, most such work has concentrated on semantic rather than syntactic issues, and where syntactic knowledge is learned, a supervised paradigm has been employed.

Thus [Brow88] learnt relationships between pairs of English and French sentences, [Mage91] worked with a treebase, and most of the connectionist work has employed the supervised backpropogation network. Even some of the work using self-organizing connectionist networks has involved a degree of supervision, e.g. [Rume86] used paired past-tense and present-tense verb forms.

In fact, the only truly *unsupervised* learning of linguistic structure, apart from [Powe89,91,92], has been the [Ritt90] work, using Kohonen nets on speech code vectors. However, the experiments in [Powe89] were also only successful when given a highly constrained input corpus, although it has since been possible to modify it to work on an arbitary corpus. On the other hand, [Gold67] proved that unsupervised learning of arbitrary language of any of the standard superfinite classes...is impossible...

In fact, we claim that it is possible to learn a natural language grammar under such conditions. We note, in particular, that in the kind of natural environment we are proposing, there is the expectation that linguistic or other motor activity by the learner will elicit responses which will provide clues to the success of his communication acts. On the other hand, psycholinguists are generally agreed that children rarely receive overt correction of their language errors, and do not appear to effect the requisite changes to their linguistic rules when they are corrected, although the evidence may be somewhat overstated.

Review of Published Experiments

The experiments of [Powe89] included two batteries of unsupervised experiments, based respectively on statistical and connectionist correlation techniques. The statistical program was fairly complex. It hypothesized unary and binary context free rules of the form:NT1 <-- NT2. NT1 <-- NT2, NT3. where Non-Terminal NT2 or NT3 could be identical with NT1, allowing recursion. All possible rules were hypothesized, and a likelihood associated with them derived from the likelihood of the components NTs. The likelihood of the non-terminals was incremented whenever it could be incorporated in a parse.

The connectionist program was much simpler, being based on a network incorporating temporal decay and delay. If a neuron fired it would decay gradually, allowing a Hebbian modification of the synapse connecting it to any neuron stimulated by the next word. Two layers of neurons were simulated, one representing words (Terminals) and the other classes (Non-Terminals).

Neither formulation worked particularly well given unrestricted text, but they were trained on children's readers in English and French and worked better, and they were trained with highly simplified artificial corpora and worked quite well. (Examples of program, data and output are given in [Powe89].)

Interestingly, when consistent rules emerged, the rules were not what I, or conventional grammar, expected. However, those with a more psycholinguistic perspective on the problem were more enthusiastic. In particular, the first class to emerge (in both systems) was the class of punctuation. The next to emerge was the class of determiners, and next the class of nouns (and adjectives). These are not unexpected classes, and it is apparent that the most 'closed' classes were being detected first. Moreover, the rules emerged in a 'bottom up' way with sentence punctuation grouping with the following article as a unit, and that aggregate grouping with the noun into a kind of subject class and they are consistent with psycholinguistic evidence that readers pay more attention to the beginnings of sentences (see [Powe89].

Insights and Experiments on Closed Classes

Linguists distinguish between Open Classes: those classes like Noun and Verb which have virtually unlimited membership (as new members can be coined); and Closed Classes: those classes like Articles, Prepositions and Pronouns, as well as Prefixes and Suffixes, which have closed membership (as those who have tried to introduce gender neutral pronouns like 'hem' have discovered). The closed class words are moreover the most frequent words in any language. Thus, 'the' alone accounts for up to 11% of a typical written English corpus (but the precise frequency is register and thus corpus dependent).

I have suggested that there is, in fact, a spectrum of closed to open classes, which is essentially a matter of class size. However, there is a complication. We often like to subdivide or subcategorize these classes in various ways. In fact, most statistical experiments on corpora have concentrated on finding

subcategorizations of the nouns and verbs (the misnomer ‘semantic’ is often used), using them to improve the probability of choosing the correct gloss in Speech Recognition or Machine Translation.

In the context of language acquisition, Psycholinguists have also postulated the existence of a pivot class of words which children use in the two word sentence stage of development. Although this has fallen out of favour, it remains true that children at this stage seem to distinguish between a closed set of words which they use as operators in conjunction with a larger set of words representing objects. The closed class is a mixed bag grammatically, hence the proposal that children are using it as a grammatical category in a child grammar. For example, the first word in each of the following child sentences is a typical ‘pivot’: big dog. want ball. allgone milk. allgone daddy.

Another significant hypothesis made in [Powe91] was that the fact that both punctuation and articles emerge first as closed classes in [Powe89], but are relative latecomers in language acquisition, indicated that these words have a prosodic function, which is like a syntactic species of deixis. In the same way that words like ‘here’ and ‘there’ direct our attention to things in the physical scene, punctuation (or prosody) and determiners direct our attention to things in the linguistic scene.

We don’t really expect that when we say ‘This is a ball’ or ‘There is the ball’ that the child will repeat ‘this’ or ‘there’ or ‘is’ or ‘a’ or ‘the’ since they have little semantic function, but rather a ‘pointing’ function of one sort or another (not all languages have words corresponding to all of these, but may rely on prosody to convey some of these distinctions). The word that is important for the child is ‘ball’. But the child hears and recognizes these other words, and already at just 4 days old can distinguish different languages - perhaps because they are characterized by these very frequent segments/words like ‘th/the’.

By the time we note that these words are frequent and highly correlated with their contexts, we have two mechanisms which allow automatic filtering of arbitrary text into the kind of simple phrases which the [Powe89] programs handled so effectively. Finding closed classes thus seems to be highly useful.

[Powe91] made the further observation that since closed classes are so easily detectable at the word level, in such an unsupervised way, perhaps they are also detectable at other linguistic levels. Since the approach used had been inspired by the Tagmemic/Phonemic analytic method, a new study was carried out at the character level and the vowels emerged as the first closed class (or if included in the corpus the punctuation again). The rules that emerged [Powe92] tended to follow sonority lines, and capture the nucleus of a syllable or morph, and then extend out to capture closed class words and affixes, and then noun phrases (without embedded clauses or multiplication of adjectives) - thus at the highest level duplicating the results of [Powe89]. Note too that this work was replicated in both Dutch and English.

The method used in [Powe91,92] was not so much statistical as distributional. In Phonemics, what is important is whether certain juxtapositions of phones can or cannot occur in the language. Thus we looked at the set of contexts (of around 5 characters) for a candidate unit - which could be ambiguous as to whether it is one character, like ‘t’, or two, like ‘th’, or three, like ‘sch’. We then looked for sets of units that occured with exactly the same contextual distribution.

We note, however two points of detail. First, we allowed for a discrepancy of up to two contexts, and eliminated the rarest sequences, to allow for errors, foreign words, etc.) Second, [Powe91] used frequency to find the ‘seed’ of the first closed class and then used context matching to ‘grow’ this into a class. More recent experiments [Finc93] have demonstrated that context-vectors alone can be used to identify the vowel class and has argued against using frequency to decide if concepts are similar. It has also been demonstrated that with more sophisticated statistical techniques single units of context suffice.

In fact frequency alone can also identify this robost class! However, we have not been able to reproduce the hierarchical classification from characters with either mechanism alone, although [Finc93] succeeds in learning a two level grammar starting from word level. Note however that in the [Powe92] experiments, frequency serves to identify the class which will next be ‘assimilated’ as a new unit and cohesiveness is not adequate as distances to growing classes are typically such that several are growing in parallel (alternately). [Finc93] manually sets a cut off level and extracts multiple classes in one go.

We later performed further experiments in which all punctuation and word cues (spaces) were filtered out and found that we were again able to reliably discover the vowel class and thus the syllable nucleus [Schi94], and we achieved this in German and French as well. Note that using a German bible as a corpus there was a tendency for ‘p’ to look a bit like a vowel as ‘sprach’ (’said’) and its derivatives occur so frequently and provides a C_C context. The letter ‘y’ in the German Eberfelder bible is always a vowel, and was grouped with the vowels along with the umlauted variants. In English, it is truly ambiguous. We also ran experiments with all characters and found groups of digits, bracketing punctuations, sentence level punctuation, etc. Note that the liquids, ‘r’ and ‘l’, also group robustly. Pilot experiments on some phonetic corpora and on speech code vectors, also show promising results.

Metrics and transformations

The focus of this is limited to the character level, and aims to systematically evaluate techniques at this level, selecting for one language and linguistic task. In the future we will treat this as hypotheses for universal mechanisms, and as with our previous work examine how well they perform on other languages and levels. In particular, we wish to explore alternatives to some expensive and linguistically and neurologically less plausible mechanisms proposed by [Finc93] who reports "best results" in all his character, phonetic, syntactic and lexical experiments using Spearman Rank Correlation, but he does not attempt the hierarchical learning of [Powe92], and he did not include binary distributional vectors (as used in [Powe91,92]) in his study, and he does not present quantitative results or comparisons. We are undertaking a study comparing binary with logarithmic, linear, reciprocal and ranked probability vectors, using a variety of metrics and normalization techniques, and present preliminary results.

Preliminary results and conclusions

Some examples of our results and different methods for displaying the results are given in the figures, and we proceed directly to discussion and conclusions from the data.

The advantage of the Spearman Rank Correlation over direct use of frequencies or (after normalizing) probabilities may not so much relate to corpus variation (as suggested by [Finc93, p94]) as to the weight attached to the different contexts. Since Zipf's law suggests that rank and frequency have an approximately reciprocal relationship, the reciprocal of frequency may be a cheap substitute and has not been investigated before. This can be taken as *the distortion hypothesis* - that the advantages of rank stem from a distortion of the context space: The ranks for the most frequent contexts are by definition very small, o(1), and are evenly spaced, and the reciprocal achieves a similar result. Because Zipf's law needs to be modified by a small exponent which is slightly different for each corpus, and can be as low as 0.5, the square root of reciprocal is also performed. Conversely the logarithm of the frequency provides an even more radical distortion and a reduction of dynamic range. The binary occurence data can be regarded as the extreme reduction of dynamic range, to just one bit, and thus fits in here too. A logical extension of this idea, is to reduce the precision of the data, or to simulate ranks most closely, to truncate the reciprocals of the (conditional) probabilities to an integer, but this avenue has not been explored.

An alternative, *the statistical hypothesis*, which is the basis put forward by [Finc93], is that rank will tend to be more stable with a small corpus and in the presence of noise. the ranks for the most frequent contexts are by definition very small, o(1), and are evenly spaced. This means they will be dominated by the sheer numbers of less frequent concepts and their increased sensitivity to noise and sample variation. This is not the case when using raw probabilities, but will be the case to an extent when using reciprocal probabilities. On the other hand there is no such bias when using binary distributions, and logarithm of probability (information) will rank in between.

Our preliminary results indicate that ranks do not work consistently well with the Euclidian metric which underlies the correlation technique of Spearman, but that the simple Manhattan metric is even better, and that in this case, binary and information (logarithm of probability) work as well or better than straight probability or ranked probability. On the other hand, rank is the most consistent performer.

An example of raw figures from the different metrics is given in Fig. 1 for two corpora of relatively small scale. One corpus is a mock medical report on a male contraceptive device, the umbrelly, which accounts for the tendency of y to emerge as a vowel. The other is the whole of Alice in Wonderland which though larger tends to lose u. Column 1 is the metric times 10000 (in brackets we have contexts/occurences for each concept along with the total number of Ngrams stored for the corpus -— they are not the same as we threshold out contexts which aren't repeated). The output is preceded by a count of lines, words and characters for each text. This Manhattan binary is the underlying metric used in [Powe91,92]. We use grep for statistics involving vowels as the formation of the class of consonants is interleaved.

We also provide a 2-dimensional map and represent the reciprocals here in such a way that the average distance between vectors is the same as in the hard to visualize multi-dimensional context space. We used steepest descent based on the total variance (error) between true and displayed distance [Samm67].

The traditional way of presenting such information is a dendrogram [Finc93] , and this interleaving can be seen there from the fact that dendrograms don't consist of monotonely increasing distances to the forming concept. However producing dendrograms requires some additional assumptions about how things group: how far apart are clusters. We address this further below and present some dendrograms in Fig. 2. Often frequency (or perhaps its inverse correlate, class size) is used as a weighting factor in this clustering process, but in Fig. 2 we follow [Finc93] in using an unweighted average of the distances between vectors. Another approach would be to define an average or centroid for each cluster.

With separate contexts (Fig. 2) rather than a combined context (Fig. 1), even using the same small single character contexts, all metrics tend to do better except for the binary occurence data where the situation is reversed. The combined context would be expected to occur with a probability which is the product of the probability of the individual contexts, and they are individually much rarer, thus there is too much noise for even rank correlation for these small corpora. But the occurence information is stronger and a few characterizing templates can define a class accurately using just binary occurence information.

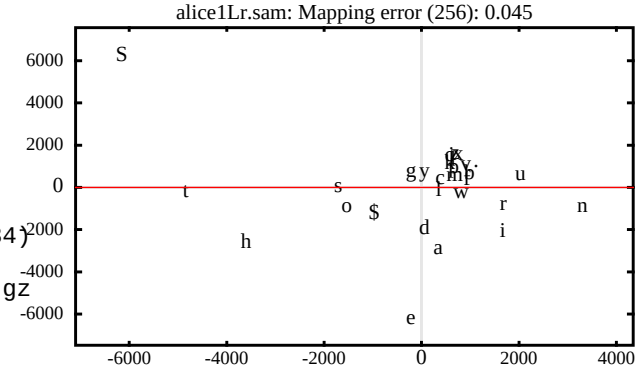
Context and Concept Size and Hyperclasses

When a child, or a linguist, is exposed to a language and seeks to learn or analyze it, he is not told, 'These are the significant sounds and this is how they vary with context.' Rather he has to learn this from the context. Working with English characters is rather similar. The character 't' represents a certain sound, but the combination 'th' represents a couple of completely different sounds (cf 'thy thigh tie').

The linguist seeks to use knowledge of words, like 'thy thigh tie die', to distinguish whether related sounds are actually different or not, and in these cases will conclude there are four different sounds involved as four (or more) distinct meanings are involved and are consistently associated with the particular phonetic sequences, which are there determined to be phonemic.

Because we cannot assume that the representation we are given involves the requisite segmentation into units for our classification, [Powe91] introduced a fuzziness into the algorithm which allow both

```
 3610   26457  148482 alice
------> alice.1:1:1.manhattan.binary.gz
6363 -> i  a  (6740/7511+7896/8787:6284)
7311 -> o  a  (7286/8141+7896/8787:6284)
7394 -> o  e  (7286/8141+12323/13569:6284)
7526 -> o  i  (7286/8141+6740/7511:6284)
7605 -> e  i  (12323/13569+6740/7511:6284)
7746 -> e  a  (12323/13569+7896/8787:6284)
7993 -> _  e  (24566/26458+12323/13569:6284)
8172 -> o  u  (7286/8141+2970/3468:6284)
------> alice.1:1:1.manhattan.information.gz
6721 -> i  a  (6740/7511+7896/8787:6284)
7518 -> o  e  (7286/8141+12323/13569:6284)
7562 -> o  a  (7286/8141+7896/8787:6284)
7574 -> e  i  (12323/13569+6740/7511:6284)
7601 -> o  i  (7286/8141+6740/7511:6284)
7775 -> e  a  (12323/13569+7896/8787:6284)
8064 -> _  e  (24566/26458+12323/13569:6284)
8083 -> o  u  (7286/8141+2970/3468:6284)
------> alice.1:1:1.manhattan.probability.gz
7141 -> i  a  (6740/7511+7896/8787:6284)
8264 -> ,  e  (2295/2418+12323/13569:6284)
8300 -> o  i  (7286/8141+6740/7511:6284)
8317 -> o  e  (7286/8141+12323/13569:6284)
8391 -> o  a  (7286/8141+7896/8787:6284)
8464 -> ‘  i  (956/1108+6740/7511:6284)
8491 -> ,  o  (2295/2418+7286/8141:6284)
8551 -> y  e  (1956/2264+12323/13569:6284)
8553 -> :  o  (179/233+7286/8141:6284)
8553 -> o  ;  (7286/8141+158/194:6284)
8646 -> ‘  a  (956/1108+7896/8787:6284)
8655 -> e  i  (12323/13569+6740/7511:6284)
8664 -> _  e  (24566/26458+12323/13569:6284)
8687 -> w  u  (2433/2674+2970/3468:6284)
------> alice.1:1:1.manhattan.ranked.gz
6894 -> i  a  (6740/7511+7896/8787:6284)
7624 -> o  e  (7286/8141+12323/13569:6284)
7889 -> o  i  (7286/8141+6740/7511:6284)
7916 -> e  i  (12323/13569+6740/7511:6284)
8040 -> o  a  (7286/8141+7896/8787:6284)
8071 -> _  e  (24566/26458+12323/13569:6284)
8235 -> e  a  (12323/13569+7896/8787:6284)
8379 -> s  e  (5703/6495+12323/13569:6284)
8438 -> t  e  (9928/10684+12323/13569:6284)
8509 -> n  u  (6491/7013+2970/3468:6284)
```




```
 55     436    2821 umbrelly
------>
umbrelly.1:1:1.manhattan.binary.gz
8571 -> o  a  (93/143+111/161:1152)
8666 -> a  i  (111/161+116/163:1152)
8800 -> a  e  (111/161+206/302:1152)
8888 -> u  o  (31/63+93/143:1152)
------>
umbrelly.1:1:1.manhattan.information.gz
8613 -> o  a  (93/143+111/161:1152)
8786 -> a  e  (111/161+206/302:1152)
8850 -> a  i  (111/161+116/163:1152)
8862 -> u  o  (31/63+93/143:1152)
8996 -> e  i  (206/302+116/163:1152)
------>
umbrelly.1:1:1.manhattan.probability.gz
8135 -> a  i  (111/161+116/163:1152)
8537 -> u  i  (31/63+116/163:1152)
8779 -> o  a  (93/143+111/161:1152)
8846 -> o  .  (93/143+26/32:1152)
8918 -> u  a  (31/63+111/161:1152)
8932 -> y  e  (26/43+206/302:1152)
------>
umbrelly.1:1:1.manhattan.ranked.gz
8223 -> a  i  (111/161+116/163:1152)
8783 -> o  a  (93/143+111/161:1152)
8792 -> y  e  (26/43+206/302:1152)
8867 -> o  e  (93/143+206/302:1152)
8936 -> u  o  (31/63+93/143:1152)
8965 -> a  s  (111/161+106/144:1152)
8988 -> e  i  (206/302+116/163:1152)
```

Figure 1. *Results indicating corruption of vowel class using ranks and probabilities. Distances are computed using the L1 metric. The concept and left_right context are limited to single characters. The Sammon map is computed using the L2 metric displaying results for the reciprocal of frequency (odds).*

concepts and contexts to be multicharacter. This introduces a degree of representation independence. At this point it appears that only I and my collaborators are pursuing this facet, and indeed I have been criticized for this added complexity [Schi93,Finc93].

Nevertheless, it is clear that the task of dealing with 'phonetic' transcriptions (technically they are phonemic as there are phonetic distinctions that are not captured in the notation) is simpler as we have perfect knowledge of the segmentation. The same applies at the level of dealing with words (although where we put spaces in English does not always correspond with any consistent grammatical definition of word, and the same is true for other languages).

What the critics have missed, however, is the enormous payoff of this approach. The classes contain not just single characters but pairs of characters [Powe91,92] and not just digraph consonants (like 'th' and 'qu') but vowels (like 'ou' and 'ai'). Furthermore, if we replace a class by a new non-terminal, this non-terminal becomes a unit available for participation in this same process. This is the way vowels and liquids combine together into syllable nuclei etc. It is also the way, at higher levels, we end up with hyperclasses of phrase structure rules. In fact, for a language like German, with composite left_right contexts, it doesn't matter where the filler is 'ö' or 'o"' and there is absolutely no difference in the way these corpora will be treated (and the umlaut vowels do group correctly). More interesting will be to see how they compare in the 'oe' alternate convention, which is not so clear cut as 'e' has is itself graphemic.

[Finc93,p88] argues that rather than allowing multicharacter contexts, Ngrams, we should use the concatenation of the bigram tables at different offsets, viz the character directly to the left, two to the left, etc. separately, and similarly for the right context. This clearly allows operation with smaller corpora (as the larger the Ngram the larger the corpus required to develop reliable statistics). Although the multiple bigram approach is ubiquitous in corpus linguistics, the Ngram approach has served well in the [Powe91,92] experiments, and the multibigram doesn't match with phonological practice where the longest possible matching contexts are sought (namely identical word environments). In particular, [Powe91] showed, counterintuitively, that right context is more useful than left, but that the combination of the separate distances did not improve, and sometimes weakened, right context alone. Using vectors whose contexts are bisymmetric, that is Ngram rather than multibigram, did seem to produce more stable classes.

Moreover, the kind of multicharacter segments we are looking for have higher frequencies than single characters like 'x', 'z' and 'q'. The noisy combinations which have no significance tend to be eliminated by simple application of a threshold. But for very frequent characters, we will also find that combinations with other characters are above threshold. These combinations will then get assessed for distributional similarity to other segments and will be rejected if they do not belong to any hyperclass. A class of singleton segments which is meaningful will clearly be preferred to a meaningless class involving spurious multicharacter segments and little disturbance of the classes has been observed due to considering multicharacter segments. Occasionally something like 'or' will tend to be added to the vowels under certain choices of metric and vector, preempting the next pass when the class of vowels, V, is permitted to combine with the liquids to produce a new class of syllabus nucleii. However, using the binary distributional analysis of [Powe91], three of the four distance measures we employed never lead to this problem (but one did see rare occurences).

Nonetheless, we are currently studying the effect of using multibigram contexts rather than Ngram contexts in combination with the study on vector types and metrics. Note that there will be an error introduced by not distinguishing the context th_C from the contexts t__C and h_C. In particular t__C will now tend to propose a hyperclass involving segments 'he' and 'ea' (due to words like 'them' and 'team'). Thus we expect that whilst it will increase reliability when multicharacter segments are excluded (that is we assume that they don't exist and that 'h' is some kind of modifier like 'r'), it will decrease reliability when multicharacter segments are proposed.

Preliminary results strongly confirm the advantage of the Ngram over the multibigram, especially in relation to the combined use of left and right context. In the multibigram classes, 'r' tends to join the vowels very early, and in some metrics 's' and 't' join before 'u'.

Future Work

The current studies discussed above are concentrating on consolidating the existing work by exploring alternate formulations of the problem. In fact, I have always emphasized that what is significant in this work is the paradigm rather than the particular algorithmic details, vectors, measures and metrics used - indeed similar results have been achieved using totally different approaches, at the level of classification. However, up to now only the distributional approach of [Powe91,92] has been used successfully in

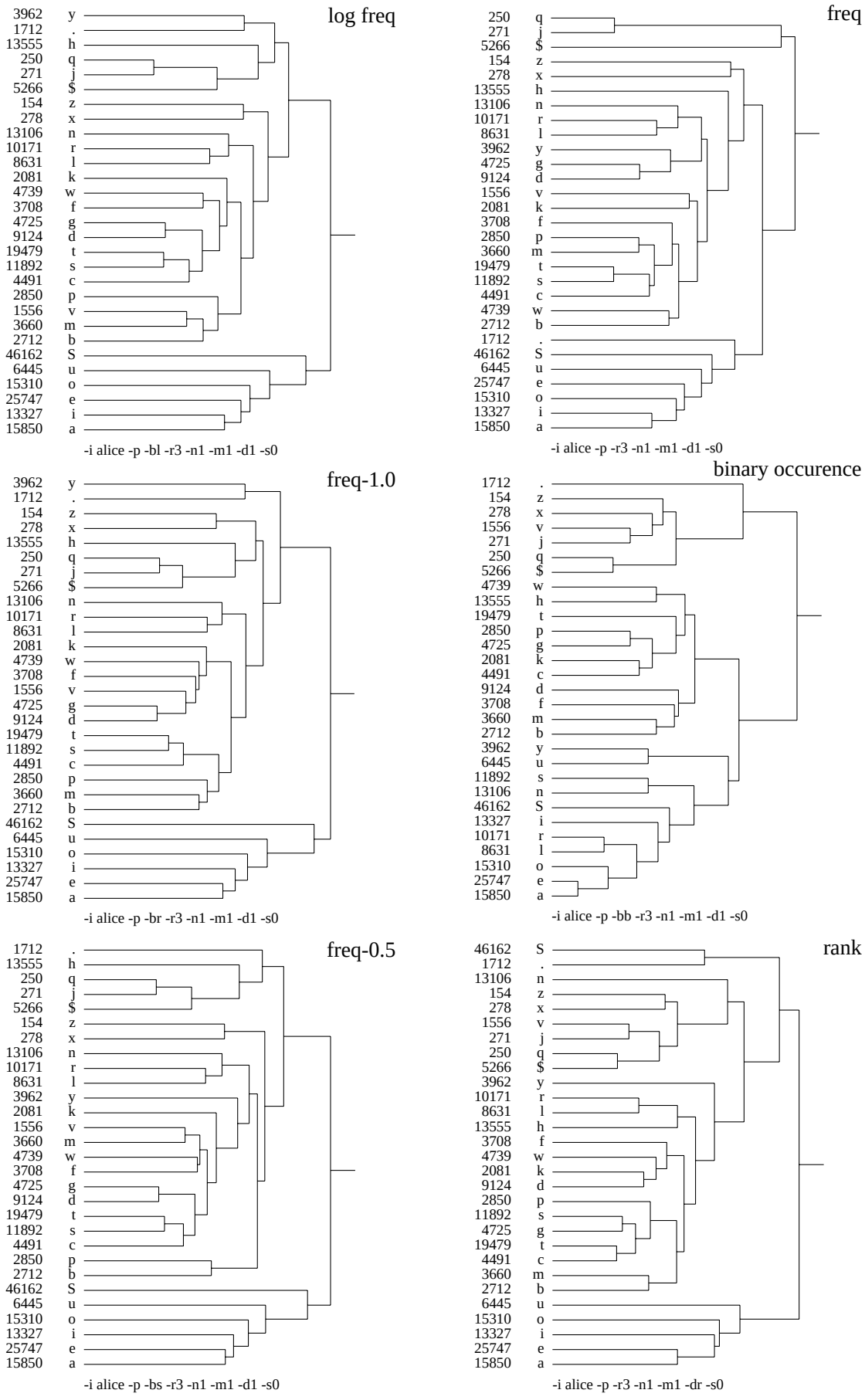


Figure 2. *Dendrograms of graphemes using ranks and functions of probabilities. Distances are computed using the L1 metric. The concept and* separate *left and right contexts are single characters.*

learning hierarchical structures - although some alternate approaches have been tried unsuccessfully. Furthermore the space of choices of metric, normalization, transformation has not been explored. We have introduced some goodness measures for comparing the approaches. Figure 3 illustrates the ratio of minimum distance to a non-member over maximum distance within the target class. Further measures are used to take into account choice of clustering mechanism as well.

References

[Brow88] Brown, P.J., Cocke, S., Della Pietra, V.J., Della Pietra, F. Jelinek, R. L., Mercer R.L.& Roossin, P., 'A Statistical Approach to Language Translation', *Proceedings of COLING 88*, 71-76

[Finc93] Steven Finch, 'Finding Structure in Language', *PhD Thesis*, University of Edinburgh UK

[Gold67] E. Mark Gold, "Language Identification in the Limit",*Information and Control* **10**, pp 447-474.

[Mage91] David M. Magerman and Mitchell P. Marcus, 'Distituent Parsing and Grammar Induction; The Automatic Acquisition of Linguistic Structurefrom Large Corpora.' pp. 122-125 in Powers & Reeker, *Procs AAAI Spring symposium on MLNLO*, DFKI, Univ. Kaiserslautern FRG.

[Pike47] Pike, K.L. *Phonemics: A Technique for Reducing Languages to Writing*, U. Michigan Press

[Pike77] Pike, K.L & Pike E.G. *Grammatical Analysis*, SIL and U. Texax, Arlington

[Powe89] David M. W. Powers and Christopher Turk, *Machine Learning of Natural Language*, Research Monograph, Springer-Verlag, 1989.

[Powe91] David M. W. Powers, 'How far can self-organization go? Results in unsupervised language learning.' pp. 131-137 in Powers & Reeker, *Procs AAAI MLNLO*, DFKI, Univ. Kaiserslautern FRG.

[Powe92] David M. W. Powers, 'On the Significance of Closed Classes and Boundary Conditions: Experiments in Lexical and Syntactic Learning', pp. 245-266 in Walter Daelemans and David M. W. Powers, Eds, *Background and Experiments in Machine Learning of Natural Language: First SHOE Workshop*, ITK Proceedings 92/1, Tilburg University NL.

[Ritt 90] Ritter, H. Kohonen, T. 'Learning Semantotopic Maps from Context', *Proceedings of the IJCNN*, Washington, 1990.

[Rume86] James L. McClelland,David E. Rummelhart, "On Learning the Past Tenses of English Verbs'' pp216-272 in *Parallel Distributed Processing*, Vol2: Psychological and Biological Models, James L. McClelland, David E. Rummelhart and the PDP Research Group, MIT Press 1986.

[Samm67] John W Sammon, Jr., 'A Nonlinear Mapping for Data Structure Analysis', *IEEE C*-**18** #5, 401-409

[vdM91] Christoph von der Malsburg, 'Self-Organization of Orientation Selective Cells in the Striate Cortex', Kybernetik 14:85-100.

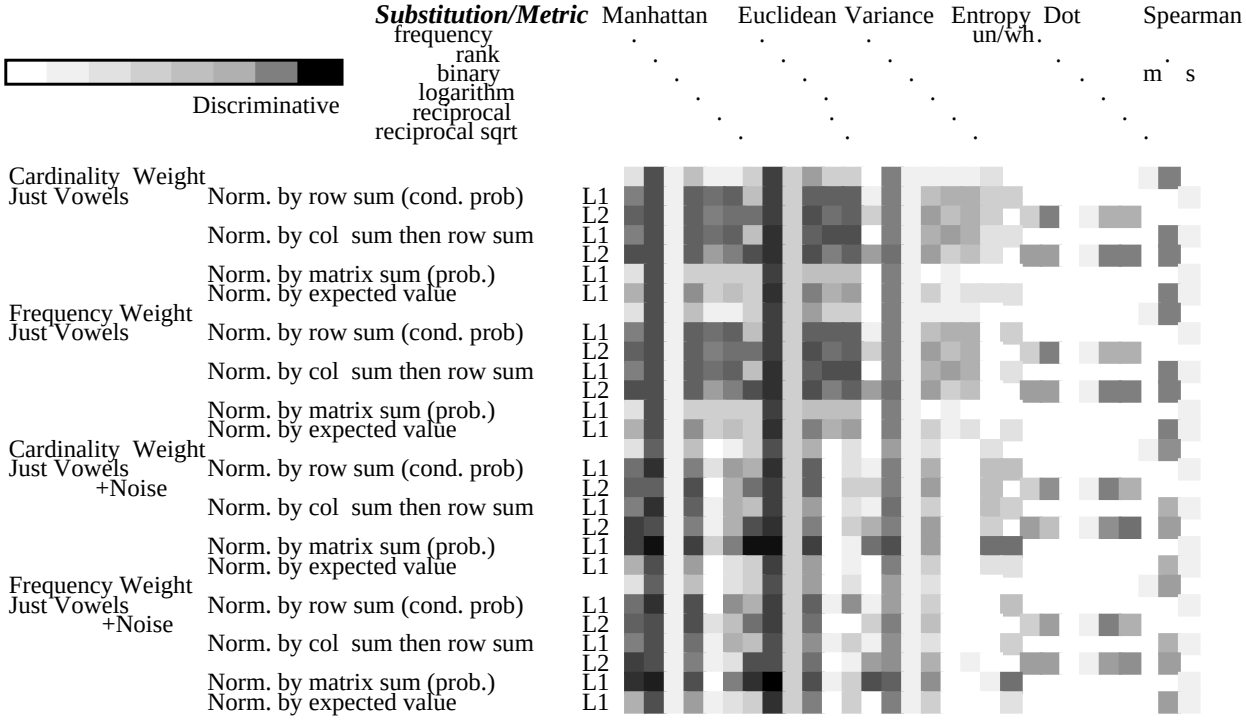


Figure 3. *Comparison matrix for dendrograms of graphemes using ranks and functions of probabilities. The density represents the ratio of the 'no man's land' around a class to the diameter of the class.*