Fachbereich Informatik
Universität Kaiserslautern
Postfach 3049
D-6750 Kaiserslautern

SEKI - REPORT

Experiments in Unsupervised Language Learning

David M. W. Powers

SEKI Report SR-90-20

Experiments in Unsupervised Language Learning

David M. W. Powers[1]
Universität Kaiserslautern
D-6750 KAISERSLAUTERN
WEST GERMANY

powers@informatik.uni-kl.de
Tel: (+49-631) 205-3449
Fax: (+49-631) 205-3200

Abstract

There are a number of debates in linguistic, psycholinguistic and neurolinguistic circles which have relevance to research on machine learning of natural language. Some of these concern where language lies on the spectrum between innate and learnt; how much can be learnt in the absence of semantics; how much can be achieved by neural self-organization without multi-layer back-propogation; and how important negative information is to language learning.

The computational research presented in this paper places a point of reference on each of these spectra, and indeed suggests that they are not independent.

We present some computational experiments and results, and propose ideas towards a theory of language learning. More importantly we pose some traditional questions in a new light and suggest new avenues of research for the traditional cognitive science disciplines as well as modern computational linguistics.

The concrete experiments presented use statistical techniques for lexical learning and were inspired by earlier experiments using statistical and neural techniques for syntactic learning and speech recognition. The interrelationships and significance of all of these experiments are discussed.

---

[1] The author is supported under ESPRIT BRA 3012: COMPULOG.

Introduction

[Gold67] and [Mins69] produced results which demonstrated limitations on the possibility of learning. These were based on certain assumptions about the learning mechanisms and the problem domain, and were in various respects both intended and construed as criticisms of current approaches and claims. In the first case, [Gold67] showed that context free languages couldn't be learned without either a teacher or a critic. In the second case, [Mins69] showed that a class of (visually presented) group invariant relations could not be recognized by Perceptrons.

Since then, the more powerful PDP (Parallel Distributed Processes) approach popularized by [Rume86] (and subsequent publications from the same group) has demonstrated overwhelmingly that useful learning (*inter alii* in the language and vision domains) can be done with neural nets. In a less focussed way, MLNL (Machine Learning of Natural Language) has also found renewed vigour [Lang87; Powe91].

But there are still things our machines can't yet do. And there are still things our machines can't ever do. The results hold. But there are things we, that is humans and other organisms, can do. And there are language, vision and speech features that earlier statistical and neural models did learn [Koho84,89,90; Powe83,89; Ritt89]. The trick is to characterize these accurately and discover appropriate mechanisms – whether they be the *natural* mechanisms, just *similarly* effective mechanisms, or *better* mechanisms.

In [Powe83,89] one of several experimental language learning programs used self-organizing neural network techniques to learn word classes and syntactic rules in a total absence of critical input. There was simply multiple exposure to a set of legal phrases, with no teacher supplying anomalous input in the sense of [Gold67]. Nonetheless, the system managed to learn the word

classes correctly, as well as grammatical rules which, if not actually those the grammarians discovered, are nonetheless effective. Similar results were achieved in a statistical program applied to the same data. The neural program was shorter. The statistical program faster.

[Koho90; Ritt89] independently showed that neural and comination statistical/neural self-organization techniques can learn word classes (but apparently *not* syntactic rules) of similar complexity in a different domain – again in the absence of critical input. (Similar techniques were applied by [Koho84] to mapping Finnish and Japanese phonemes – viz. achieving the feature/phone to phoneme classification.)

What is interesting is not just what was learned in terms of word classes, but what was learned first and why these particular rules were learned. It turns out that the most closed classes were learned first. These then seemed to act as pointers to the more open word classes they were associated with. This paper proposes that these results can give us insights as to why closed class words, such as articles, occur at all, how they are learned, and why they are not used early but are recognized. It also extends the experiments below the word level to see if there are closed classes there.

None of these previous reports or reviews has fully considered the broader computational, linguistic and psycholinguistic significance of these particular results (although [Powe91] does point to most of the issues involved). Here we consider this significance in several respects: in relation to closed classes, in relation to symbolic properties of connectionist systems, in relation to the weak form of learning used, and in relation to more accurate characterization of natural language.

Therefore, we will first summarize the methodology and results of the "noun phrase" experiments of [Powe84] and the "sentence" experiments of [Ritt89;

Koho90], we then address some of the issues to which they are relevant and introduce some hypotheses to be tested. We finally present a computational experiment using similar techniques in the new, sub word-level, domain of classification of letters/phonemes into the classes from which syllables and words are composed, giving our procedures, results and conclusions.

Review of Syntactic Learning Experiments

We do not wish to review statistical, neural or syntactic learning generally, but to take up certain experiments from [Powe89] and [Koho90], and compare the application of similar techniques in one of the domains that bridges the gap. As mentioned in the introduction, the pedigree of such work extends back beyond the criticisms of [Gold67] and [Mins69] and are reviewed and represented adequately elsewhere (see, in addition, [Lang87; Powe91] for pointers).

The experiments we wish to review were presented in the context of noun phrases and filtered sentences; the classes categorized and grammatical rules learnt were discovered with two different mechanisms, neither of which required critical input.

We imagine that the computational model represents a child at the beginning of the stage where he learns some nouns and verbs and their meanings and that he is trying to make sense at the same time of the images he is faced with. We further suppose that there are prosodic and syntactic features which tend to highlight the significant words, e.g. that they occur stressed in phrase final position. We hypothesize further that what is far beyond the child's competence and far from these significant positions is filtered out, and that conversely the child focuses on what is close to or within his competence.

We actually make no use of these assumptions other than to provide some justification for the type of dataset used for the learning experiments, which we

present in figures 1 and 2 in the form used in the simulations of [Powe89] and [Koho90] resp.

```
the cat. #
a dog. #
my dog? #
this mat! #
...
```

Fig. 1. *Example dataset à la [Powe89].*

In the original experiments the '#' of Fig. 1 had some 'monitoring' significance. It also serves as a reminder of the elision. The prosody of speech is hypothesized to have some correspondence to the punctuation symbols used in these text experiments.

```
Mary likes meat
Jim speaks well
Mary likes Jim
Jim eats often
...
```

Fig. 2. *Example dataset à la [Koho90].*

Note that both of these datasets can be regarded as sets of "three word sentences" representing utterances from which the uninteresting parts have been filtered according to different theories, or different applications of a general theory.

A first criticism can already be mentioned here: results with the omitted words included are *not* presented. Although the preliminary results from experiments with more complex data were (as could be expected) more complex and less conclusive, they would be interesting to see, and should give an idea of the degree of reliance placed on the above-mentioned assumptions. (A listing of one of the actual neural programs used is however presented in [Powe89], allowing the possibility of repetition or extension of the experiment.)

It should be noted too that the learning, particularly for the (pure) neural simulations, is very slow. For example, the "semantic map" of [Koho90: Fig.12] resulted from "2000 presentations of word-context-pairs derived from

10 000 random sentences of the kind shown". (It is therefore very time-consuming and unrewarding to explore the more unlikely directions!)

Statistical Psycholinguistic Model

The first model [Powe83,89] makes use of an additional psycholinguistic hypothesis. It uses the *Magical number seven plus or minus two* of [Mill56] to constrain the number of partial parse fragments (trees) kept around on *tags* and available for correlation. Unlike some of the earlier models, it then not only turns collocations of words into hypotheses of rules, but collocations of tags.

A second technique, also motivated by psycholinguistic considerations, is used to consolidate rules, in an induction step, bring together into the same hypothesized class words with collate similarly, viz. with the same words or classes. Thus classes are formed initially as small consistent cosets of words.

A thresholding step is used before rules are considered ready for *production* use – again a psycholinguistic hypothesis lies behind this terminolgy. It is proposed that the unthresholded grammar can play a role in guiding the recognition process in terms of indicating the likely class of a word, but that there is an implicit or explicit partitioning into *recognition* and *production* grammars mediated, in part, by some sort of threshold.

We present in Fig. 3 only a sample thresholded, consolidated grammar to give the flavour of the results.

```
      Sense Class Thresh-Set                    Sense Class Thresh-Set

                                          class(lang, 16, [a, the,...])
formula(lang, 24, [[17, 10]])             class(lang, 10, [rat, cat,...])
formula(lang, 17, [[12, 16]])             class(lang, 12, ['.','?','!'])
```

Fig. 3. *Sample output from [Powe89a].*

The first observation to be made (to an extent observable in the structure of the rules) is that the first class learnt is the *punctuation/prosody*. Next come the *articles* and finally the *nouns*. The significant aspect is that the most *closed* (or

smallest) classes are learnt first and that these act as pointers (in the rules) to the more significant *contentive* and *open* classes.

Self-Organizing Neural Net

The above experiment was duplicated [Powe83,89] with a self-organizing model inspired by the visual application of such a neural net by [Mals73], but based in some respects on the model of [Klop82]. Interestingly, this program did not make use of the magical number seven directly, but a similar effect result from the *decay* model used. Once a neuron had fired it decayed over a period of time allowing for the possibility of it interacting with the neurons firing as a result of subsequent "words".

The results of this experiment were comparable with the statistical version, and a relationship between neurons and classes, synapses and grammatical rules was apparent in the comparison of the results.

The experiments of [Ritt89; Koho90] used a similar model applied to their dataset. For efficiency they turned to a hybrid statistical/neural approach in which they first preprocessed the data to produce an "average context" for each word – an average of all code vectors of predecessor-successor-pairs surrounding the given word. Note that this windowing is very sensitive to the omitted words, but could be justified on the basis that these words really represent the phrases of which those words are the nucleus.

The methodology of [Ritt89; Koho90] is explicitly exploiting the contextual similarity of items. The important feature is that the context is consistently dominant and recognizable in the learning process, and thus words may be classified by the contexts they occur in, and that then the classification of words together allows unification of contexts and consequent strengthening of the context consistency.

In these experiments the context taken was the pair of "words" preceding and succeeding the word in focus. In the experiments of [Powe84,89] the context was determined by the *decay* mechanisms or the *tag* mechanism. In recent experiments based on the paradigm of [Ritt89], similar results have been produced with "contextual sensitivity" being provided by the addition of recurrence between layers [Scho91].

Hypotheses

The experimental perspective taken here is concerned with understanding the nature of language learning enough to implement useful models by whatever means, whether neural or statistical, hybrid or novel. And we follow [Powe89] in recognizing the importance of contributions from Cognitive Science, and our theoretical model conforms, in the main, to the hypotheses present in Chapter 13 thereof. In particular, we recognize the importance of physiological restrictions for the determination of the nature of language, we learn language by making hypotheses which can prove useful irrespective of their validity, we envisage the negative information necessary for learning as coming from the natural restrictions of human physiology, environment and current hypotheses rather than from explicit teachers and critics.

In neural networks this type of system behaviour is called *self-organization*. In other contexts it is called *auto-correlation* or *emergence*. It is can also be seen as a consequence of fundamental principles well known in Linguistics, and indeed the foundation of Phonology (and also its generalization to Tagmemics), namely: Contrast in Identical Environments (CIE) and Contrast in Analogous Environments (CAE).

We wish to develop one hypothesis further here. It is beyond the scope of this paper to go over once more the psycholinguistic evidence reviewed in [Powe89], but we note that the experiments we reviewed in the last section are

consistent with, or at least suggestive of, the *complexity hypothesis*, *pivot grammars*, and *nucleus-margin coordination*. These suggest respectively that the simplest concepts (and by extension here, constructs and classes) are learnt first; that certain words in a *child grammar* function in a special way, as *pivots*, whilst not conforming precisely to adult *grammatical classes;* and that a *binary grammar* is evident, at many levels, in which the components differ in importance and may thus be designated as *nucleus* and *margin*.

In terms of gramatical *classes*, the natural complexity metric is the size of the class. A class that is always represented by a single exemplar, or a very small number of exemplars, but whose degree of occurence is comparable with other classes, will clearly provide a unmistakable context which can act as a boundary condition for the self-organizing process. That is, *closed* classes will act as pointers to the more *open* classes. This facilitates *focussing* on the open class "word" and hence the attachment of semantics. The broader scope and easy identification of the *open* class therefore makes it the ideal candidate to be the main information carrier, or *contentive*, as well as the syntactic *nucleus*.

Mem	(Lev)	Description
??	(0)	Several independent variables determining features
11	(1)	4 to 6 feature single phone characters
10	(2)	2 or 3 character (consonant or vowel) clusters
8	(3)	2 or 3 cluster C*V+C* syllables
7	(4)	2 or 3 syllable morphs
6	(5)	2 or 3 morph words
4	(6)	2 or 3 word phrases
3	(7)	2 or 3 phrase clauses
2	(8)	2 or 3 clause sentences
1	(9)	1 or 2 sentence (nuclear or marginal) paragraph segments
.5	(10)	2 or 3 segment paragraphs
.2	(11)	1 or 2 paragraph monologues
.1	(12)	2 or 3 monologue dialogues

Fig. 4. *Phono-morpho-phraseology.* Levels of the speech-language hierarchy, from feature level through Phonology and Morphology to Phrase Structure and Discourse Grammar are illustrated with a level number for reference and an idea of the possible variation of the number of units stored and available at that level (decreasing as complexity increases).

This process can be reflected at many levels, and is by no means limited to the speech hierarchy (Fig. 4). Similar processes were indeed first observed in vision [Mals73]. But in the context of speech, the prosodic features (including

stress, intonation, speech rate and pauses) form clear easily distinguishable classes of limited membership. This allows focussing on phonological phrases and syllables. These have a close relationship to the grammatical phrase and morph, where a similar process can identify repeated syllable/morphs as contexts which will cohere into a closed class. Similarly phrases subtended by a particular closed class can act as units in which the frequently occuring templates can provide boundary conditions for the self-organization at that level.

The experiments reported above demonstrate these effects at several different levels. Phonemes have been mapped by neural self-organization; noun phrases have had their word components classified by the same and related statistical techniques; sentences have had their phrase/word components classified similarly.

We proposed to explore one of the missing pieces from this features to sentence classification: the syllable is normally defined in terms of particular patterns (varying according to language) or consonant (C) and vowel (V) classes. The syllable and these consonant vowel classifications are missing from the above demonstrations. The consonants and vowels are determined by phonetic features, and a related prosody also helps to identify syllables. Our theory would suggest that these physiological characteristics should act as restrictions (or boundary conditions) defining logical closed classes which would be actual syntactic entities, and would thus adopt also the associated syntactic and semantic properties (*open* = *contentive* = *nucleus*).

Why should we distinguish vowel and consonant – or indeed liquids, nasals, etc? Morphophonemics dictates some constraints, but why would we expect a grammatical function? This hypothesis provides an explanation. It further leads us to predict that we should discovering such a class by application of self-organization. To be more precise, we would expect the vowels to appear as a closed class rather than the consonants, being a smaller class – although

liquids or nasals or something else could be a candidate according to size, but are excluded by their lack of primary grammatical significance. As there is not a one to one correspondence between phonemes and graphemes (characters) we allow the possibility of groups of graphemes to function as a unit, and hence the possibility that diphthongs or modified characters (e.g. +h, +r, +l, etc.) might be present.

There is also the question of how small a closed class should be – even those we have identified could conceivably be subclassified. We *need* not to introduce size as a parameter, however the *magic number seven* is again used as a memory/window constraint. The vowels happen, interestingly to fall into the *magic number seven plus or minus two* range. They may just be another addition to the catalogue of its magical properties!

Algorithm

We first note that clusters or phrases (collationally significant class constructed from lower level units) are significant to the extent that:

a. Units occur relatively frequently in conjunction with their predecessor(s);

b. Units occur relatively frequently in conjunction with their successor(s);

c. Prefixed units have a considerably modified class of successors;

d. Suffixed units have a considerably modified class of predecessors;

e. Suffixed units have an almost unmodified class of successors;

f. Prefixed units have an almost unmodified class of predecessors.

Thus, /qu/ is significant by *a*, /th/ is significant by *b*, *c* and *e*, /ck/ is significant by *b*, *c* and *d*. In the case of properties *a* and *b*, one unit acts as a good predictor for the other member(s) of the cluster. Properties *c* and *d* indicate that the cluster does not simply inherit collations but has unique characteristics. The final pair of properties are related to apparent recursion, but are more general in that they extend to cohesive constraints.

Normally the modified succ/predecessors class is a reduction which excludes those which make up the other component of the structure. It may be that the class is as it would have been without the intervention (apart from such modifiers), or that it follows the modifier, or both. Thus /t/ can be followed by [h],[r],VOWEL; /th/ can be followed by [r],VOWEL; /tr/ can be followed by VOWEL; /thr/ can be followed by VOWEL. So: /th/ is a level 2 modification, /thr/ & /tr/ are level 3 clusters.

The present algorithm looks for signs of the first of these three pairs of properties: it collects all the contexts for each character and group of character within SEVEN character strings (including word boundary and capilization codes); it then groups into classes all the common characters and character groups which occur in an identical context (left and right contexts separately), associating their sets of contextual distributions with the classes; it finally seeks to correlate similar distributions (± TWO) and allows evaluation according to either symmetric or assymetric relevance, either weighted or unweighted by the size of the class found.

SEVEN and TWO are parameters which may be varied slightly. Examples of the results and the intermediate stage associations will be presented in the next section, along with some more detail concerning the transformations at each stage. An overview is presented in Fig. 5.

```
1. Read dict & produce Context-Char sets <= SEVEN chars   - fsm
2. Convert significant sets into Cluster-Cluster pairs    - gsm
3. Group left and right sets into g & h distributions     - d?sm
4. Group complementary clusters into g & h cosets/cnt     - c?sm
 4a. Intersect gives distribution for both sides          - cism
5. Restrict all distribution size to SEVEN +/- TWO        - class?
6. Autocorrelate for subset +/- TWO of distribution       - coset?
 6a. Intersect/Union for both/either side cosets          - cosetі
7. Make best SEVEN of either Intersect or Union into classes
 7a. Make classes mutually exclusive, formulate hyperclasses
```

Fig. 5. *Outline of algorithm.* The predicate names refer to PROLOG predicates from intermediate stages which are exemplified in subsequent figures. The '?' indicates where a 'g' or 'h' is substituted for the left and right distributions respectively (and where the 'i' for the intersected distribution goes).

Results

The first stage of the processing can be viewed as the construction of a finite state machine in which each occuring string of less than SEVEN characters constitutes a state and the following character occurences define a transition possibility. This representation was used for pragmatic reasons, including efficient indexing and other uses of the structure.

```
fsm(i,p,1,296).
fsm(v,a,1,297).
fsm(th,e,2,299).
fsm(abl,e,3,301).
fsm(g,i,1,308).
fsm(ab,l,2,309).
fsm('$co',n,3,310).
fsm(ra,n,2,310).
```

Fig. 6. *Finite State Machine representation of context and next character.* '$' marks a word boundary; '^' indicates the following character was upper case. Arguments are *context, focus, length of context, number of occurences in context.*

Examples are shown in Fig. 6 of the predicate *fsm*. Another predicate *gsm* provides a view of all pairs of clusters occuring with a combined total of SEVEN characters. Then for each left cluster the distribution of right clusters associated with it by *gsm* are extracted as *dgsm* and vice-versa (*dhsm*). A sample of these distributional classes is shown in Fig. 7, and it is already apparent there that the vowels, or something closely related, are a significant class.

```
dgsm(4,189,[d,l,n,r],'$^a').
dgsm(6,385,[a,e,er,o,r,u],'$^b').
dgsm(1,36,[r],'$^be').
dgsm(5,326,[a,ar,h,l,o],'$^c').
dgsm(1,36,[r],'$^ca').
dgsm(5,198,[a,e,i,o,u],'$^d').
dgsm(1,35,[l],'$^e').
dgsm(2,61,[r,re],'$^f').
dgsm(1,20,[e],'$^fr').
dgsm(4,144,[a,e,o,r],'$^g').
dgsm(3,206,[a,e,o],'$^h').
dgsm(3,106,[a,e,o],'$^j').
```

Fig. 7. *Distribution classes subtended by a given left context (extract).* Extract is for word initial contexts from proper nouns. Arguments are *size of class, occurences of class+context, class, context.*

We now repeat the exercise with *dgsm* to group together the cosets of clusters which subtend the same distributional class, *cgsm*, and vice-versa (*chsm*). Although some small groups of very closely related clusters arise as cosets, as illustrated in Fig. 8, the sets can also often be described in terms of common initial or final segments (cp. properties *c* to *f* above). But as there are many similar distributional classes which are affected by sample error in the selection of a limited dataset as well as by memory constraints with the rejection of rare collations.

```
cgsm(1,458,[a,an,e,i,ic,o,u],[pl]).
cgsm(1,2808,[a,ar,ara,as,at,e,en,er,h,ho,i,l,la,o,ol,or,os,ost,r,re,...],['$p']).
cgsm(2,1031,[a,ar,c,'c^',e,i,o,on],['$^m','^m']).
```

Fig. 8. *Cosets of left contexts subtending the same distribution class (extract).* Arguments are *size of coset, number of occurences, distribution class of clusters, coset of subtending clusters.*

So far we have performed *Contrast in Identical Environments* (CIE) type classification, now we want to perform *Contrast in Analogous Environment* (CAE) type classification to bring together similar distribution classes and combine their cosets and assess the number of different collations and occurences for these fuzzier hypersets of distribution classes. In fact, we use the sets of known distribution classes intersected with themselves to define a kernel which must be within TWO of the size of the intersecting class. For efficiency, we use as intersecting classes only those with a size in the SEVEN±TWO range. As illustrated in Fig. 9, the vowel class emerges as one of the most important of these.

```
classg(h,['$a',a,e,i,mi,o,u],[a,e,i,o,u],['$^d','$ch','$n','^d',fl]).
classg(h,['$a',a,e,i,mi,o,u],[a,e,i,o,u],[cr]).
classg(h,['$co','$i',a,co,i,o,u],[a,co,i,o,u],[s]).
classg(h,['$co','$i',a,co,i,o,u],[a,co,i,o,u],[n]).
```

Fig. 9. *SEVEN classes (right) and close intersections with left distribution classes(extract).* Distribution classes (from either left or right context) of size SEVEN±TWO are used to find other distribution classes which are similar in that the intersection with that SEVEN class differs by no more than TWO from the SEVEN class. Arguments are *source of selecting SEVEN class, SEVEN class, intersection with current (left) distribution class, coset of distribution class.*

At this point, we combine the information from left and right distributions and compute statistics based on the size of the common and total cosets of the SEVEN classes, or the number of actual occurences of subtended collations. On all four metrics, the vowels emerge as the most well defined class – with a significant lead over the runner up in second place, as shown with best seven scores for two of the metrics in Fig. 10.

```
coseti(28,84,4,12,[a,e,ea,i,in,o,u],[d,n,s,t],['$l',b,c,d,h,l,n,p,r,s,st,t]).
coseti(28,112,4,16,[c,f,g,p,s,t,v],[a,e,i,o],['$a','$re',^,a,al,an,e,en,er,i,...]).
coseti(30,144,5,24,[c,d,g,l,s,t],[a,ar,e,l,o],['$a',^,a,an,ar,e,en,er,i,in,l,...]).
coseti(30,168,5,28,[a,e,i,o,u,y],[b,c,m,p,s],['$h','$m','$s','$t',^,an,b,c,...]).
coseti(48,156,8,26,[a,e,er,o,r,u],[b,c,e,f,g,i,n,t],['$^b','$f','$p','$t',...]).
coseti(49,196,7,28,[a,e,i,o,r,ra,u],[b,c,d,f,g,r,t],['$b','$c','$d','$g',...]).
coseti(85,385,17,77,[a,e,i,o,u],[b,c,ch,d,e,f,g,l,ll,...],['$^d','$b','$c','...]).
```

Fig. 10a. *Cosets of SEVEN classes of either context sorted by occurence in intersection (extract).* Arguments are *occurences of intersection coset, occurences of union coset, size of intersection coset, size of union coset, SEVEN class, intersection coset, union coset.*

```
coseti(30,168,5,28,[a,e,i,o,u,y],[b,c,m,p,s],['$h','$m','$s','$t',^,an,b,c,...]).
coseti(49,196,7,28,[a,e,i,o,r,ra,u],[b,c,d,f,g,r,t],['$b','$c','$d','$g',...]).
coseti(3,96,1,32,[a,e,o],[y],['$^g','$^h','$^j','$^p','$^s','$cr','^g','^h',...]).
coseti(16,184,4,46,[a,e,o,u],[i,ll,mp,ri],['$^b','$^d','$ch','$l','$m','$n',...]).
coseti(15,245,3,49,[a,e,i,o,r],[ch,t,th],['$b','$c','$d','$f','$g','$p',...]).
coseti(16,232,4,58,[a,e,i,o],[k,sp,u,v],['$^l','$^m','$^n','$^r','$br',...]).
coseti(85,385,17,77,[a,e,i,o,u],[b,c,ch,d,e,f,g,l,ll,...],['$^d','$b','$c',...]).
```

Fig. 10b. *Cosets of SEVEN classes of either context sorted by size of union (extract).* Arguments are *occurences of intersection coset, occurences of union coset, size of intersection coset, size of union coset, SEVEN class, intersection coset, union coset.*

Conclusions

In these experiments using statistical techniques and a single exposure to each word of the Unix dictionary, the vowel class emerged first, suggesting it as a closed class. The cosets were primarily consonant clusters, suggested analogously as an open class. This confirmed a prediction that the vowel-consonant distinction was of significance in learning, that the vowels would emerge as a closed class providing a limited number of contexts, and that consonant clusters would emerge as open classes.

One surprise was that diphthongs were not represented, and indeed vowel-semivowel collations came nearer to achieving membership.

We suggest that the magic number seven plus or minus two [Mill56] should also encompass the number of the vowels. It was indeed a parameter in the analysis, and variation of this parameter did vary the precise class learnt, but the relationship has not yet been analyzed. However, its application to the size of the selected class seemed least decisive – similar results were achieved with 6±2 and 7±3 settings, for example.

The exclusion of diphthongs may also be an indicator that they are recognized as complex, at least in the orthography and under the assumptions behind this program. Recent psychological studies indicate that familiarity with written language may necessary to the (conscious) recognition of segments [Read86; Mann86]. But are diphthongs recognized as complex? Are vowels recognized as having features? Is this totally acoustic or does it have a motor component? It will be very interesting to see what results of similar experiments achieve on speech!

Although this experiment was performed using statistical techniques rather than neural networks, it was guided by previous work which achieved similar results using either or a mix, and it is expected that similar results could straightforwardly be achieved in a neural simulation.

The success of back-propogation in multi-layer neural nets has perhaps overshadowed self-organization in simpler networks, despite the impressive early low-level results; the need for semantics has perhaps overshadowed the internal consistency of grammar at the lower levels; the theoretical need for negative information from the environment has perhaps overshadowed the effective supply of criticism from boundary conditions and system restrictions; and more generally the tendency to assume that basic linguistic distinctions are innate and very closely tied to the perceptual system itself may overshadow the fact that some of these distinctions can be learnt very easily with very basic

mechanisms. These alternative perspectives are worthy of more emphasis and study.

This paper has presented some computational results and hypotheses about language learning. More importantly it poses some traditional questions in a new light and suggests new avenues of research for the traditional cognitive science disciplines.

References

[Gold67] E. M. Gold, "Language Identification in the Limit", *Information and Control* **10** 447-474 (1967).

[Koho84] T. Kohonen, K. Mäkisara, and T. Saramäki, "Phonological Maps - insightful representation of phonological features for speech recognition", *Proc. 7th Int. Conf. on Pattern Recognition* 182-185 (Montreal Canada, 1984)

[Koho89] Teuvo Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, BERLIN FRG (3rd edn, 1989)

[Koho90] Teuvo Kohonen, "The Self-Organizing Map", *Proc. of the IEEE* **78** 1464-1480

[Klop82] Klopf, A. Harry, *The Hedonistic Neuron: A Theory of Memory, Learning and Intelligence*, Hemisphere, WASHINGTON DC (1982).

[Lang87] P. Langley, "Machine Learning and Grammar Induction", Editorial to Special Issue, *Machine Learning* **2** 5-8 (1987)

[Mals73] Malsburg, C. von der, "Self-Organization of Orientation Selective Cells in the Striate Cortex", *Kybernetik* **14** 85-100 (1973).

[Mann86] Virginia A. Mann, "Phonological awareness: The role of reading experience", *Cognition* **24** 65-92 (1986).

[Mill56] George A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information", *Psychologcal Review* **63** 81-97 (1956)

[Mins69] Marvin Minsky and S. Papert, "Perceptrons", MIT Press (1969).

[Powe83] David M. W. Powers, "Neurolinguistics and Psycholinguistics as a Basis for Computer Acquisition of Natural Language," SIGART **84** 29-34 (June 1983).

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

[Powe91] David M. W. Powers, "Goals, Issues and Directions in Machine Learning of Natural Language and Ontology". Chairman's background paper, AAAI Spring Symposium on *Machine Learning of Natural Language and Ontology*, Stanford CA (March 1991).

[Read86] Charles Read, Zhang Yun-Fei, Nie Hong-Yin and Ding Bao-Qing, "The ability to manipulate speech sounds depends on knowing alphabetic writing", *Cognition* **24** 31-44 (1986).

[Ritt89] H. Ritter and T. Kohonen, "Self-Organizing Semantic Maps", *Biol. Cyb.* **61** 241-254 (1989)

[Rume86] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, MIT Press, Cambridge MA (1986).

[Scho91] J. C. Scholtes, "Learning Simple Semantics by Self-Organization", submitted to AAAI Spring Symposia on *Machine Learning of Natural Language and Ontology*.and *Connectionist Natural Language Processing*, Stanford CA (March 1991).