Unsupervised learning of natural languages

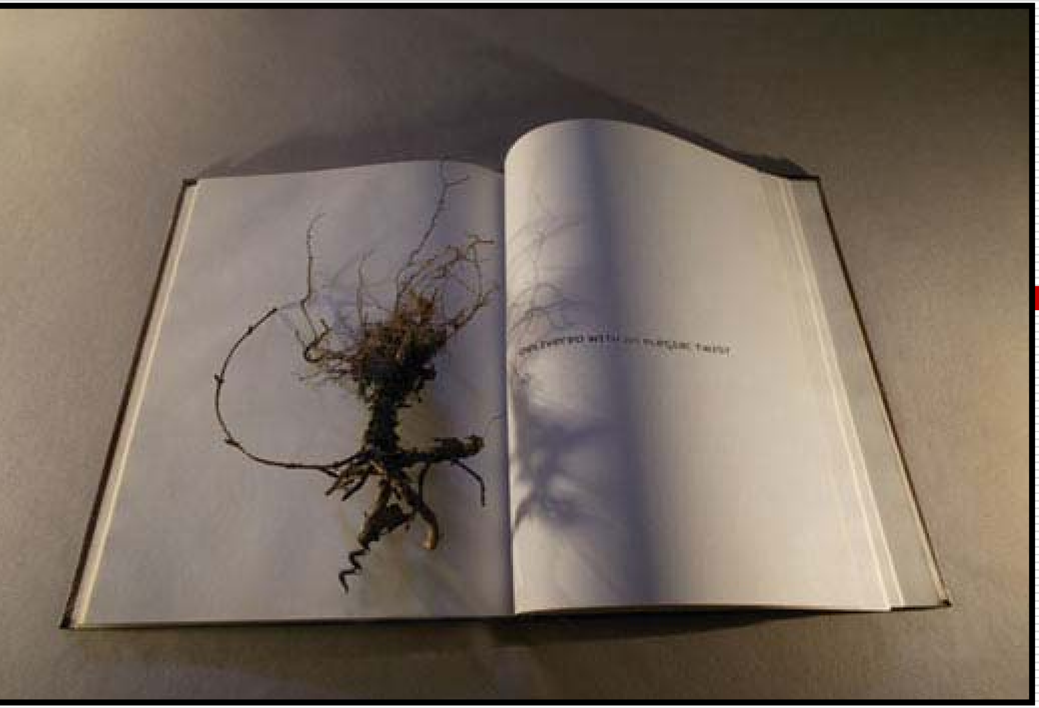
David Horn

Tel Aviv University

http://horn.tau.ac.il

in collaboration with

Zach Solan, Eytan Ruppin, Shimon Edelman



“Unsupervised Learning of Natural Languages”

Zach Solan, David Horn, Eytan Ruppin and Shimon Edelman

in *Proc. National. Academy of Sciences*. 102:11629-11634, August 16, 2005

We address the problems, fundamental to both linguistics and bioinformatics, of

- Motif extraction
- Grammar induction

i.e., inferring in an **unsupervised** manner what are the significant patterns in a text, and finding a set of rules that govern its production.

Given a corpus of strings (such as text, transcribed speech, nucleotide base pairs, amino acid sequence data, musical notation, etc.), our unsupervised algorithm **MEX** finds in it the significant motifs and **ADIOS** recursively distills from it hierarchically structured patterns via two integrated processes of segmentation and generalization.

Many types of sequential symbolic data possess structure that is (i) hierarchical, and (ii) context-sensitive.

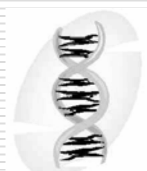
- Natural-language text:

ALICE was beginning to get very tired of sitting by her sister on the bank and of…

- Transcribed speech

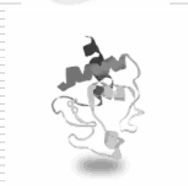
- Music

- DNA sequences



ACTTGGAATTGATCCGTATAAAT…

- Protein sequences



EFSNYKEQVAEQLIKSITQLYHD…

Toy problem:
Finding words in strings of letters

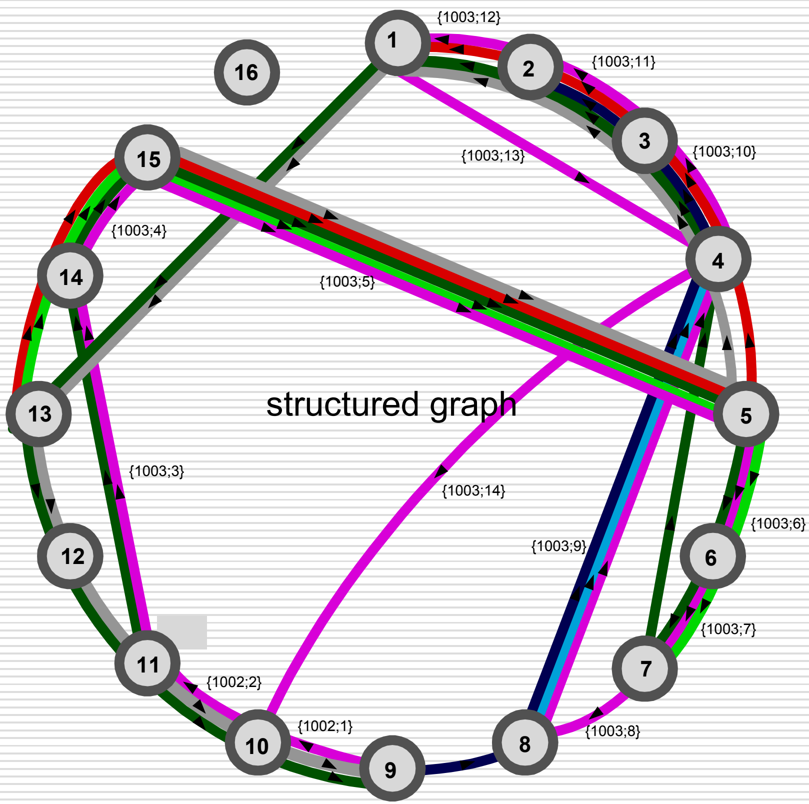
a l i c e w a s b e g i n n i n g t o g e t v e r y t i r e d o f s i
t t i n g b y h e r s i s t e r o n t h e b a n k a n d o f h a v i n
g n o t h i n g t o d o o n c e o r t w i c e s h e h a d p e e p e d
i n t o t h e b o o k h e r s i s t e r w a s r e a d i n g b u t i t
h a d n o p i c t u r e s o r c o n v e r s a t i o n s i n i t a n d
w h a t i s t h e u s e o f a b o o k t h o u g h t a l i c e w i t h
o u t p i c t u r e s o r c o n v e r s a t i o n
(A)

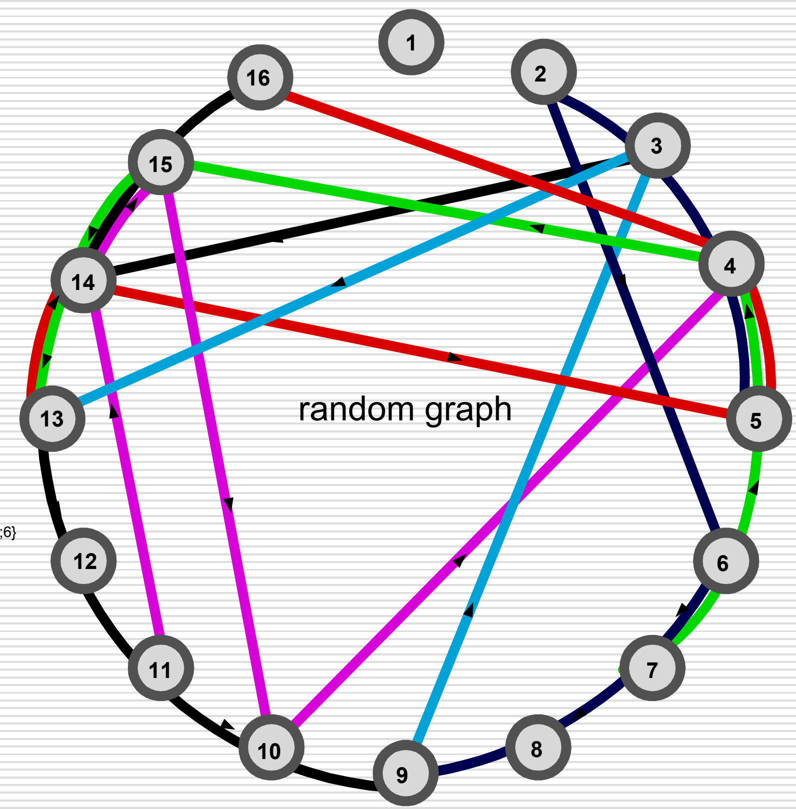
alicewas beginning toget very tiredof sitting by h ersiste r
onthebank andof having nothing todo onceortwice shehad peeped intothe
book h ersiste r was reading but ithad no picture so r conversation s
in itand w hatisthe useof abook thought alice without picture so r
conversation
(B)

MEX: motif extraction algorithm

- Create a graph whose vertices are letters
- Load all strings of text onto the graph as paths over the vertices
- Given the loaded graph consider trial-paths that coincide with original strings of text
- Use context sensitive statistics to define left- and right-moving probabilities that are used to label the beginning and end-points of motifs

Sum of paths defines a structured graph





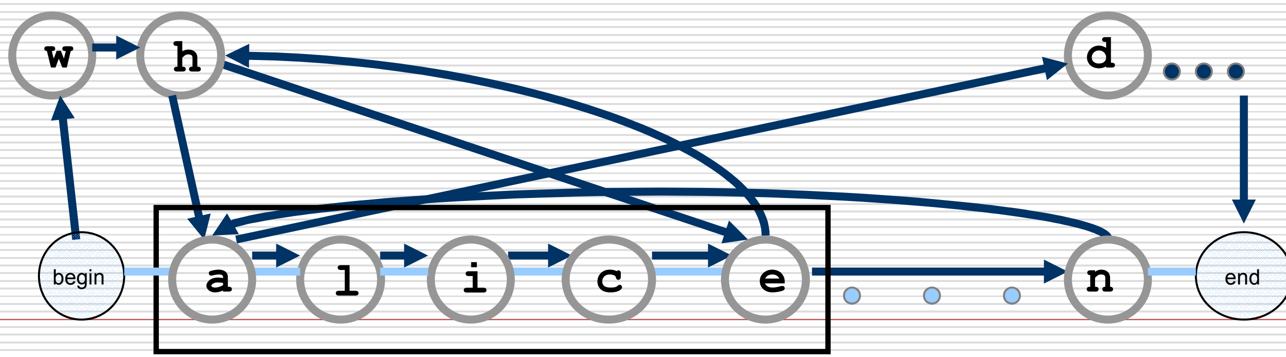
Creating the graph - cont'd

(1)

a l i c e w a s b e g i n n i n g t o g e t v e r y t i
r e d o f s i t t i n g b y h e r s i s t e r o n t h e
b a n k a n d o f c o n v e r s a t i o n

(2)

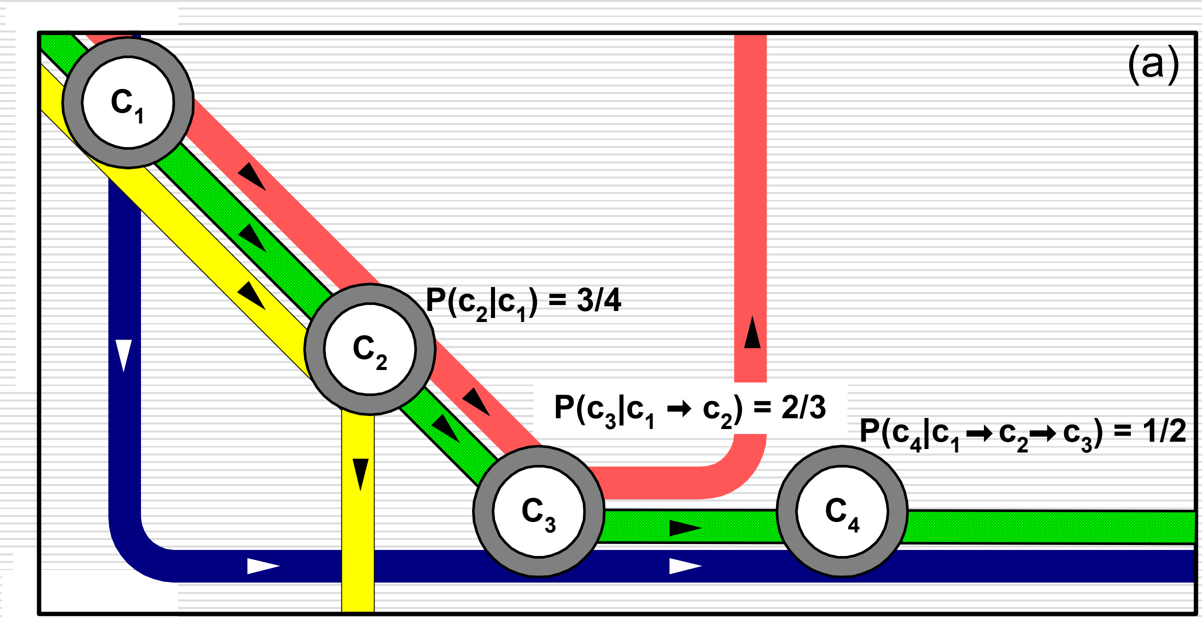
w h e n a l i c e h a d b e e n a l l t h e w a y d o w
n o n e s i d e a n d u p t h e o t h e r t r y i n g
e v e r y d o o r s h e w a l k e d s a d l y



Number of paths, L

L	a	l	i	c	e	w	a	s
a	8770							
l	1046	4704						
i	468	912	7486					
c	397	401	637	2382				
e	397	397	488	703	13545			
w	48	48	51	66	579	2671		
a	21	21	21	23	192	624	8770	
s	17	17	17	19	142	377	964	6492
b	2	2	2	2	5	10	14	63
e	2	2	2	2	4	6	9	24
g	2	2	2	2	4	5	5	10

paths allow for the definition of conditional probabilities of (almost) any order.



Probabilities are proportional to the corresponding number of paths (through-moving flux/incoming flux).

Calculating conditional probabilities

L	
a	**8770**
l	**1046**
i	**486**
c	**397**
e	**397**
w	48
a	21
s	17

P(a) = 0.08

P(l|a) = 1046/8770

P(i|al) = 486/1046

P(c|ali) = 397/486

P(e|alic) = 397/397

P(w|alice) = 48/397

P_R	
a	**0.08**
l	**0.12**
i	**0.45**
c	**0.85**
e	**1**
w	0.12
a	0.44
s	0.81

Right-moving probability

P_R	a	l	i	c	e	w	a	s
a	0.08							
l	0.12	0.043						
i	0.45	0.19	0.069					
c	0.85	0.44	0.085	0.022				
e	1	0.99	0.77	0.3	0.12			
w	0.12	0.12	0.1	0.094	0.043	0.024		
a	0.44	0.44	0.41	0.35	0.33	0.23	0.08	
s	0.81	0.81	0.81	0.83	0.74	0.6	0.11	0.059
b	0.12	0.12	0.12	0.11	0.035	0.027	0.015	0.0097
e	1	1	1	1	0.8	0.6	0.64	0.38
g	1	1	1	1	1	0.83	0.56	0.42

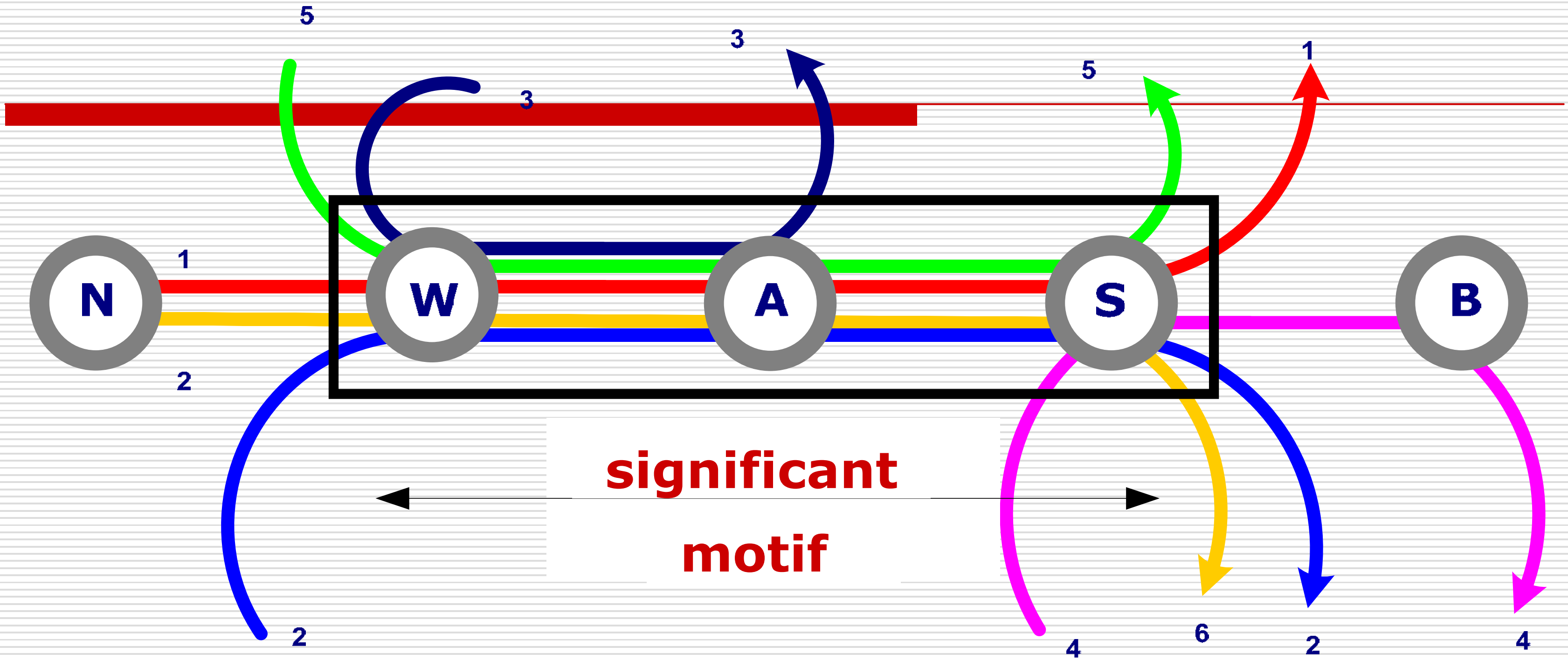
Path dependent probability matrix containing variable order Markov chains

$$M(e_1 e_2 \cdots e_k) \doteq \begin{pmatrix} p(e_1) & p(e_1|e_2) & p(e_1|e_2e_3) & \dots & p(e_1|e_2e_3\dots e_k) \\ p(e_2|e_1) & p(e_2) & p(e_2|e_3) & \dots & p(e_2|e_3e_4\dots e_k) \\ p(e_3|e_1e_2) & p(e_3|e_2) & p(e_3) & \dots & p(e_3|e_4e_5\dots e_k) \\ \vdots & \vdots & \vdots & & \vdots \\ p(e_k|e_1e_2\dots e_{k-1}) & p(e_k|e_2e_3\dots e_{k-1}) & p(e_k|e_3e_4\dots e_{k-1}) & \dots & p(e_k) \end{pmatrix}$$

P_R defined going top down ; P_L defined going bottom up

Once the graph is loaded with all data, search for patterns is carried out along trial-paths, following the paths of the data.

Searching for motifs



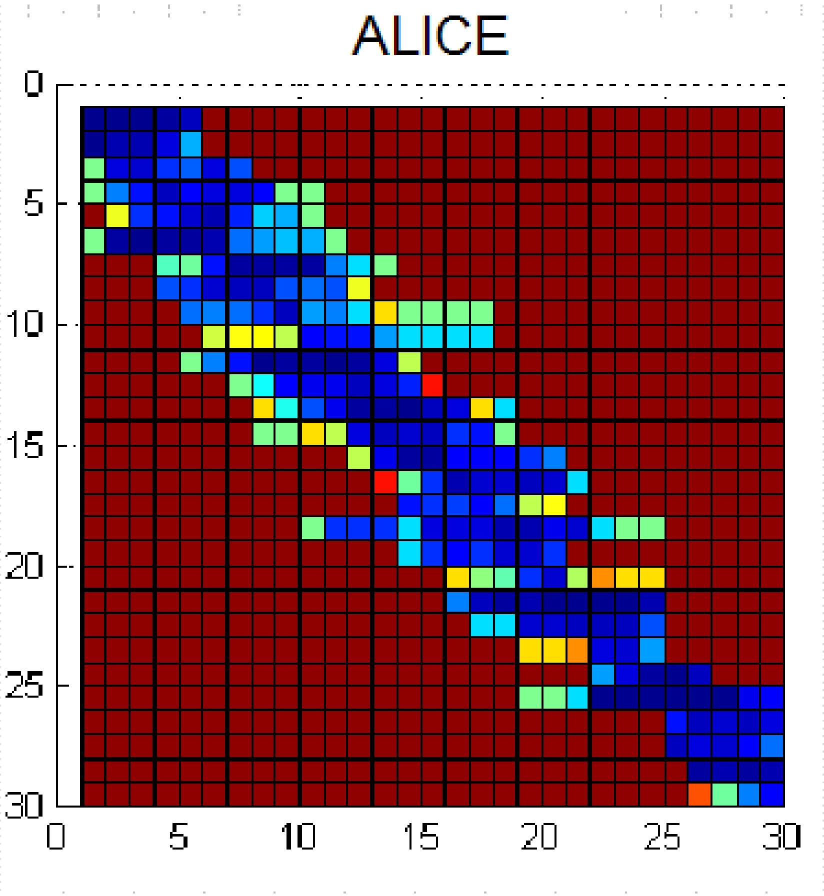
- **Mathematical formulation:**

$$M_{ij}(S)=\begin{cases} P_R(e_i,e_j) & \text{if } i>j \\ P_L(e_j,e_i) & \text{if } i<j \\ P(e_i) & \text{if } i=j \end{cases}$$

$$D_R(e_i,e_j)=\frac{P_R(e_i,e_j)}{P_R(e_i,e_{j-1})}<\eta$$

$$D_L(e_j,e_i)=\frac{P_L(e_j,e_i)}{P_L(e_{j+1},e_i)}<\eta$$

Matrix of probabilities



The MEX algorithm

Evaluate the matrix of probabilities.

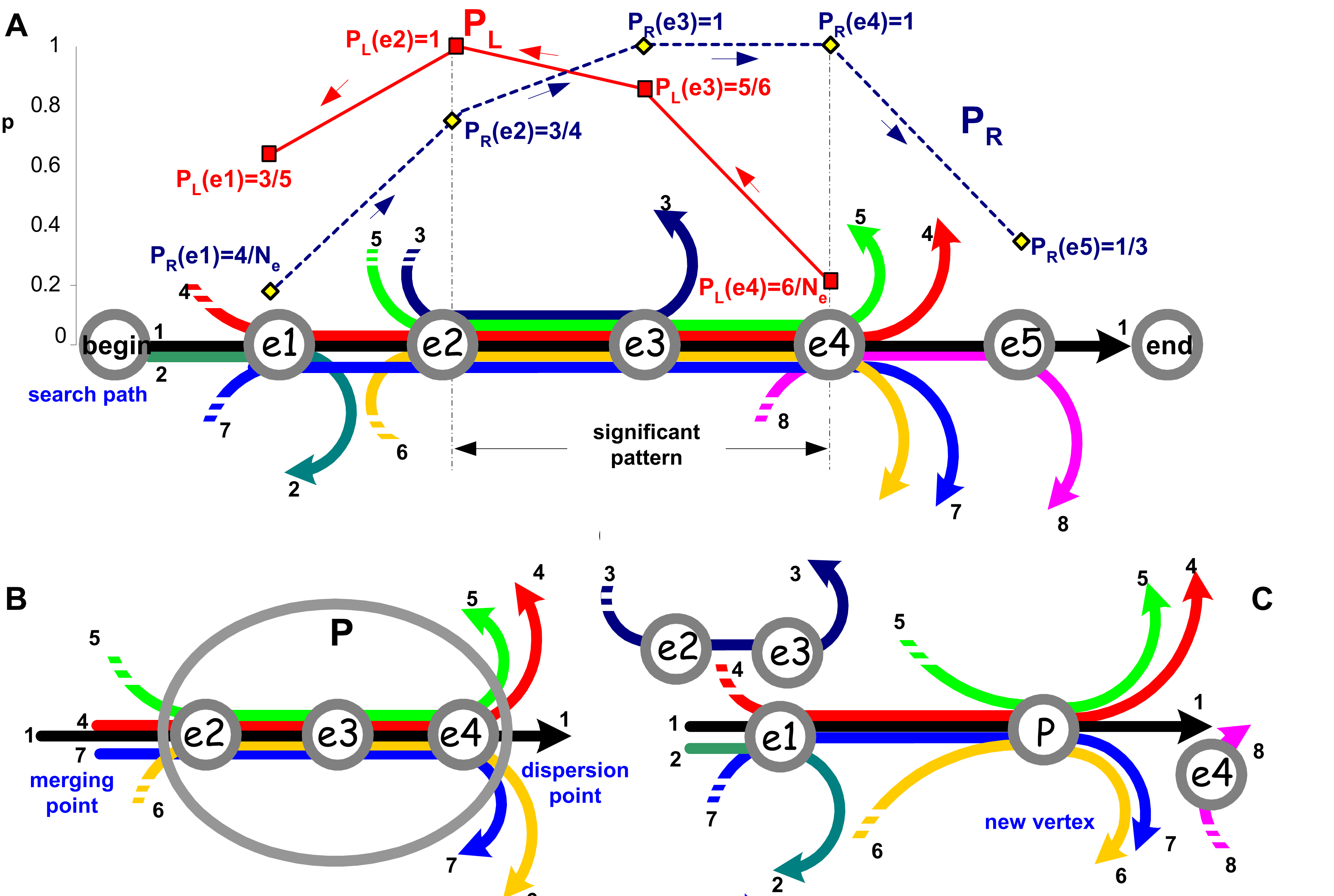
Find candidates for beginning and end-points of motifs.

Check the significance (1-α) of P_R decrease to decide on the end-point.

Rewire graph by adding the motifs as new vertices, starting with the longest and most significant motifs.

Option: Repeat with higher values of α.

The MEX (motif extraction) procedure



ALICE motifs

Motifs selected in order of

-length

-weight (significance of drop)

Shown here are results of one run over a trial-path and the beginning of the list of motifs extracted from it

	Weight	Occurrences	Length
conversation	0.98	11	11
whiterabbit	1.00	22	10
caterpillar	1.00	28	10
interrupted	0.94	7	10
procession	0.93	6	9
mockturtle	0.91	56	9
beautiful	1.00	16	8
important	0.99	11	8
continued	0.98	9	8
different	0.98	9	8
atanyrate	0.94	7	8
difficult	0.94	7	8
surprise	0.99	10	7
appeared	0.97	10	7
mushroom	0.97	8	7
thistime	0.95	19	7
suddenly	0.94	13	7
business	0.94	7	7
nonsense	0.94	7	7
morethan	0.94	6	7
remember	0.92	20	7
consider	0.91	10	7
curious	1.00	19	6
hadbeen	1.00	17	6
however	1.00	20	6
perhaps	1.00	16	6
hastily	1.00	16	6
herself	1.00	78	6
footman	1.00	14	6

The first paragraph of ALICE using MEX analysis with increasing values of α=0.001,0.01,0.1,0.5

(A) a l i c e w a s b e g i n n i n g t o g e t v e r y t i r e d o f s i t t i n g b y h e r s i s t e r o n t h e b a n k a n d o f h a v i n g n o t h i n g t o d o o n c e o r t w i c e s h e h a d p e e p e d i n t o t h e b o o k h e r s i s t e r w a s r e a d i n g b u t i t h a d n o p i c t u r e s o r c o n v e r s a t i o n s i n i t a n d w h a t i s t h e u s e o f a b o o k t h o u g h t a l i c e w i t h o u t p i c t u r e s o r c o n v e r s a t i o n

(B) alice was begin n ing toget very t i re do f sitting b y hersister onthe b an k and of ha v ing no thing to do on c eortw i ce shehad p ee p ed in tothe b ook hersister was reading but it hadno p i c t u re s or conversation s in it and what is the us e of a b ook thought alice with out p i c t u re s or conversation

(C) alice was beginning toget very tiredof sitting b y hersister onthe b an k and of ha v ing no thing to do on c eortw i ce shehad p ee p ed in tothe b ook hersister was reading but it hadno picture s or conversation s in it and what is the us e of a b ook thought alice with out picture s or conversation

(D) alice was beginning toget very tiredof sitting b y hersister onthe bank and of ha v ing nothing to do on c eortw i ce shehad p ee p ed in tothe b ook hersister was reading but it hadno picture s or conversation s in it and what is the us e of a b ook thoughtalice with out picture s or conversation

(E) alicewas beginning toget very tiredof sitting by hersister onthe bank and of having nothing to do onceortwice shehad peep ed intothe b ook hersister was reading but it hadno picture s or conversation s in it and what is theuseof ab ook thoughtalice without picture s or conversation

Application to Biology

Data: protein sequences in terms of 20 amino acids.

Example: using MEX to search for motifs in a family of 6600 enzymes, after which same motifs are used as the basis for **functional** classification with SVM.

Success measured by J=tp/(tp+fp+fn)

Vered Kunik, Zach Solan, Shimon Edelman, Eytan Ruppin and David Horn, CSB 2005.

Extracting Motifs from Enzymes

- Each enzyme sequence corresponds to a single path

>P54233 | 1.7.1.1

LLDPRDEGTADQWIPRNASMVRFTGKHPFNGEGPLPRLMHHGFITPSPLRYVRNHGPVP
KIKWDEWTVEVTGLVKRSTHFTMEKLMREFPHREFPATLVCAGNRRKEHNMVKQSIGFNWGA
AGGSTSVWRGVPLRHVLKRCGILARMKGAMYVSFEGAEDLPGGGGSKYGTSVKREMAMDPSRDI
ILAFMQNGEPLAPDHGFPVRMIIPGFIGGRMVKWLKRIVVTEHECDSHYHYKDNRVLPSHVDA
ELANDEGWWYKPEYIINELNINSVITTPCHEEILPINSWTTQMPYFIRGYAYSGGGRKVTRVEVT
LDGGGTWQVCTLDCPEKPNKYGKYWCWCFWSVEVEVLDLLGAREIAVRAWDEALNTQPEKLI
WNVMGMMNNCWFRVKTNVCRPHKGEIGIVFEHPTQPGNQSGGWMAKEKHLEKSSES

- Applying MEX to oxidoreductases
- 6602 enzyme sequences
- MEX motifs are **specific** subsequences

Enzymes Representation

- Each enzyme is represented as a 'bag of motifs'

>P54233 | 1.7.1.1
**LLDPRDEGTADQWIPRNASMVRFTGKHPFNGEGPLPRLMHHGFITPSPLRYVRNHGPVP
KIKWDEWTVEVTGLVKRSTHFTMEKLMREFPHREFPATLVCAGNRRKEHNMVKQSIGFNWGA
AGGSTSVWRGVPLRHVLKRCGILARMKGAMYVSFEGAEDLPGGGGSKYGTSVKREMAMDPSRDI
ILAFMQNGEPLAPDHGFPVRMIIPGFIGGRMVKWLKRIVVTEHECDSHYHYKDNRVLPSHVDA
ELANDEGWWYKPEYIINELNINSVITTPCHEEILPINSWTTQMPYFIRGYAYSGGGRKVTRVEVT
LDGGGTWQVCTLDCPEKPNKYGKYWCWCFWSVEVEVLDLLGAREIAVRAWDEALNTQPEKLI
WNVMGMMNNCWFRVKTNVCRPHKGEIGIVFEHPTQPGNQSGGWMAKEKHLEKSSES**

↓

>P54233 | 1.7.1.1
**RDEGTAD, TGKHPFN, LMHHGFITP, YVRNHGPVP, WTVEVTG, PDHGFP
YHYKDN, KVTRVE, YGKYWCW, MGMMNNCWF**

- These 1222 MEX motifs cover 3739 enzymes

Enzyme Function

- The **functionality** of an enzyme is determined according to its **EC number**
- **EC number: $n_1.n_2.n_3.n_4$** (a unique identifier)
- **Classification Hierarchy** [Webb, 1992]
  - **n_1**: **class**
  - **$n_1.n_2$**: **sub-class / 2nd level**
  - **$n_1.n_2.n_3$**: **sub-subclass / 3rd level**
  - **$n_1.n_2.n_3.n_4$**: **precise enzymatic activity**

An example:

- EC 1 .12 . 1 . n_4

oxidoreductases

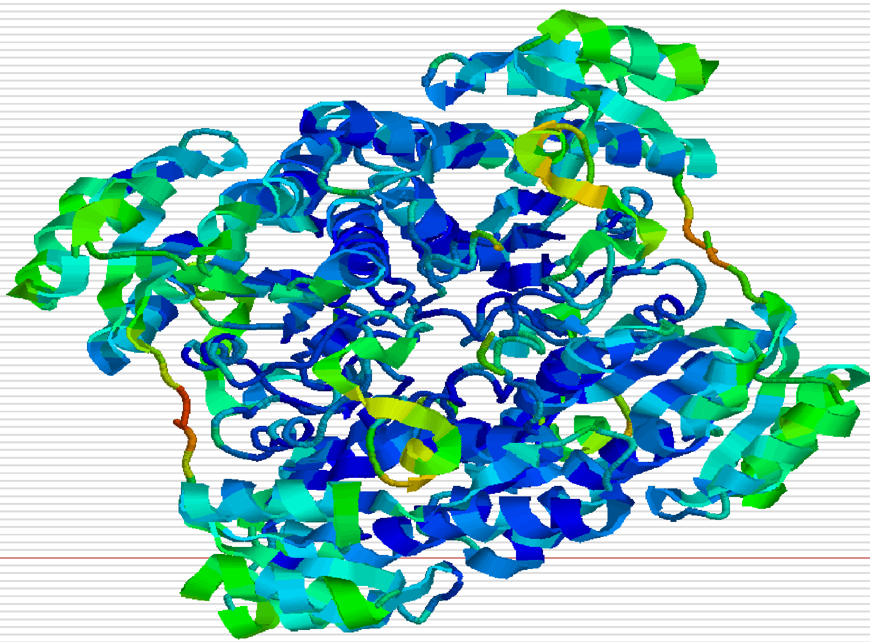
hydrogen as electron donors

NAD+ / NADP+ as electron acceptors

NAD+ oxidoreductase

EC 1.12.1. 2

$$H_2 + NAD^+ = H^+ + NADH$$



The MEX method

- **SVM classifier input:**

O17433	1148	262	463	610	7987	1627	260
P19992	124	7290	27	111	3706	18128	3432
Q01284	6652	198	1489	710	425	64	55
Q12723	693	145	7290	3712	65	543	522
P14060	455	2664	848	55	128	256	74
Q60555	7290	3712	65	543	522	6748	7159

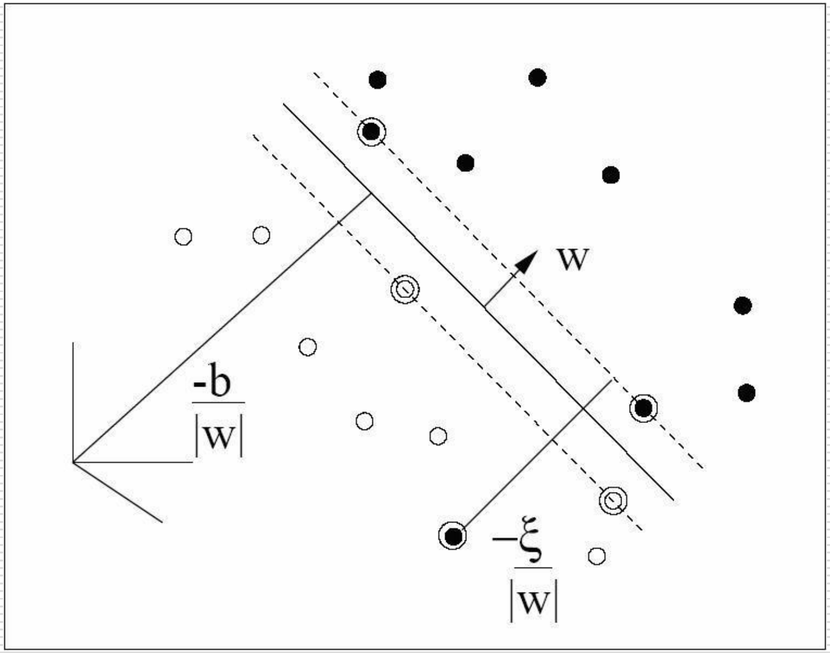
Classification Tasks:

- **16** 2^{nd} level subclasses
- **32** 3^{rd} level sub-subclasses

Basic notions in linear SVM

- Given a set of data points **x** that correspond to classes y=(1,-1), i.e. given pairs of {**x**,y}, we ask for their best linear separator:
- Find **w** such that **w.x**+b>1 defines the class y=1, while **w.x**+b<-1 defines the class y=-1.
- M=2/|**w**| is the optimal margin of separability.
- **w** can be expressed in terms of the linear superposition of a few of the data-points **x**.
- These few lucky ones are called support-vectors. They lie on the two separating planes.
- The SVM method can handle outliers.
- SVM-light is available on the internet. It chooses the best parameters by itself.

SVM



Maximize:

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

subject to:

$$0 \leq \alpha_i \leq C,$$

$$\sum_i \alpha_i y_i = 0.$$

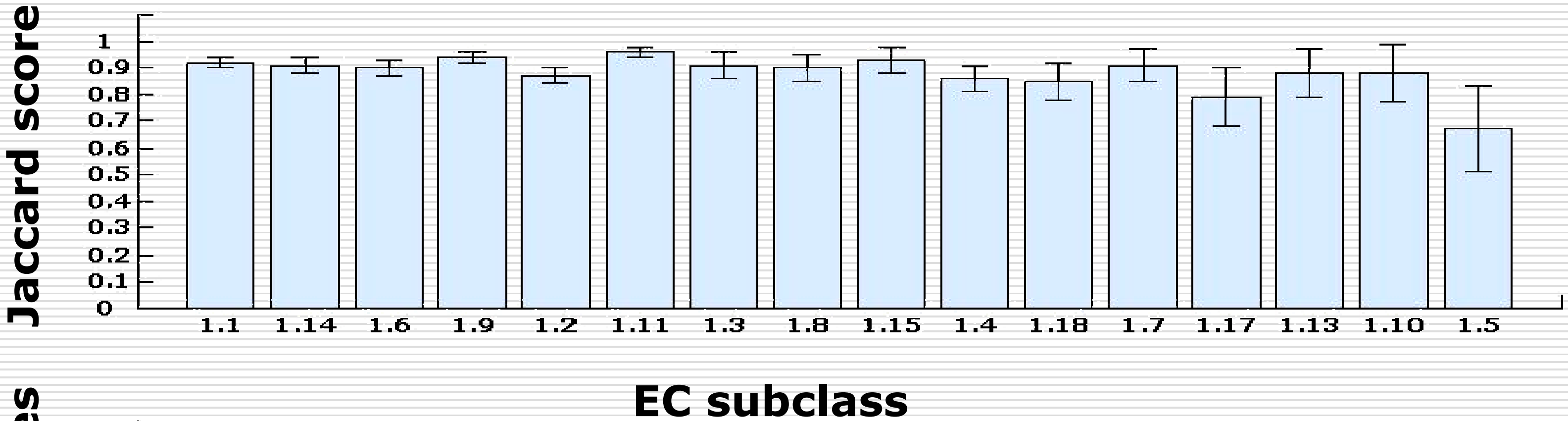
The solution is given by

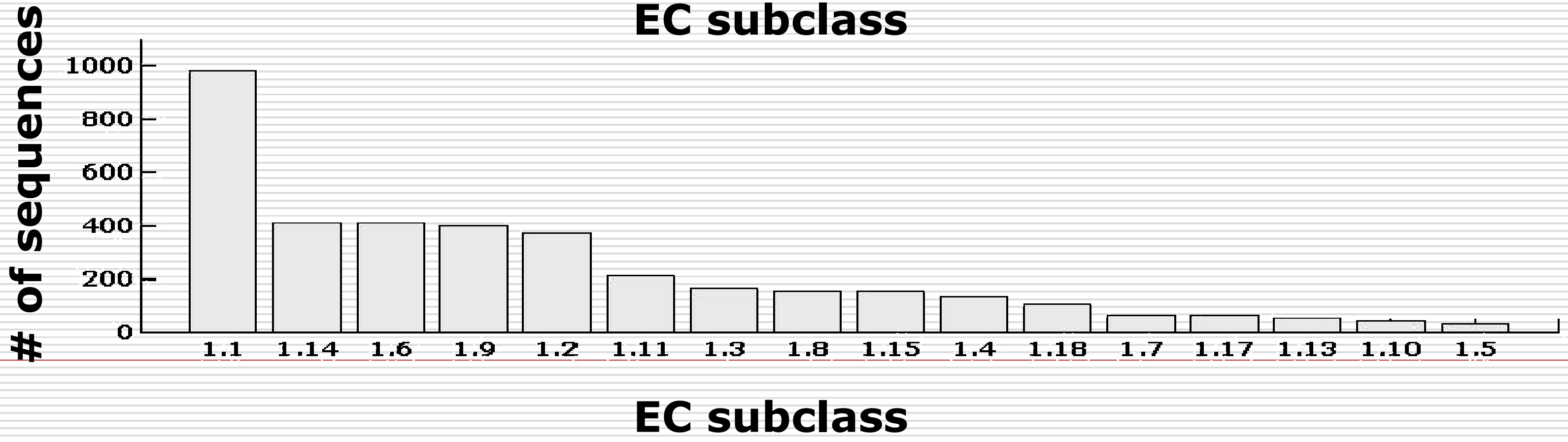
$$\mathbf{w} = \sum_{i=1}^{N_S} \alpha_i y_i \mathbf{x}_i.$$

Results

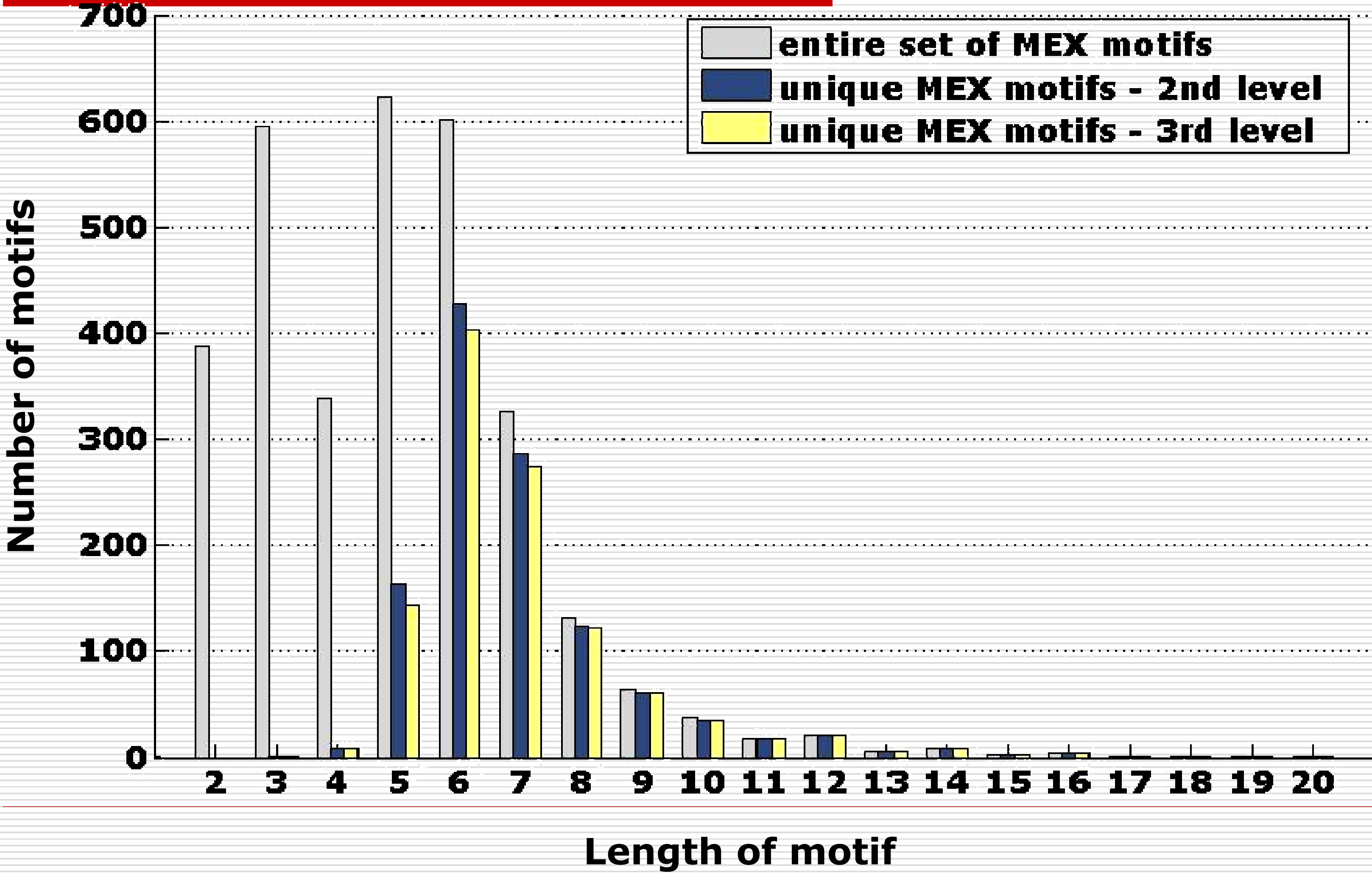
- **Average Jaccard scores:**
  - **2nd level: 0.88 ± 0.06**
  - **3rd level: 0.84 ± 0.09**

2nd level results





Results of the Analysis – cont'd



ADIOS (Automatic DIstillation Of Structure)

- Representation of a corpus (of sentences) as paths over a graph whose vertices are lexical elements (words)
- Motif Extraction (MEX) procedure for establishing new vertices thus progressively redefining the graph in an *unsupervised* fashion
- Recursive Generalization

Purpose

Goal: achieve an integrated understanding of acquisition and representation of linguistic structures that is

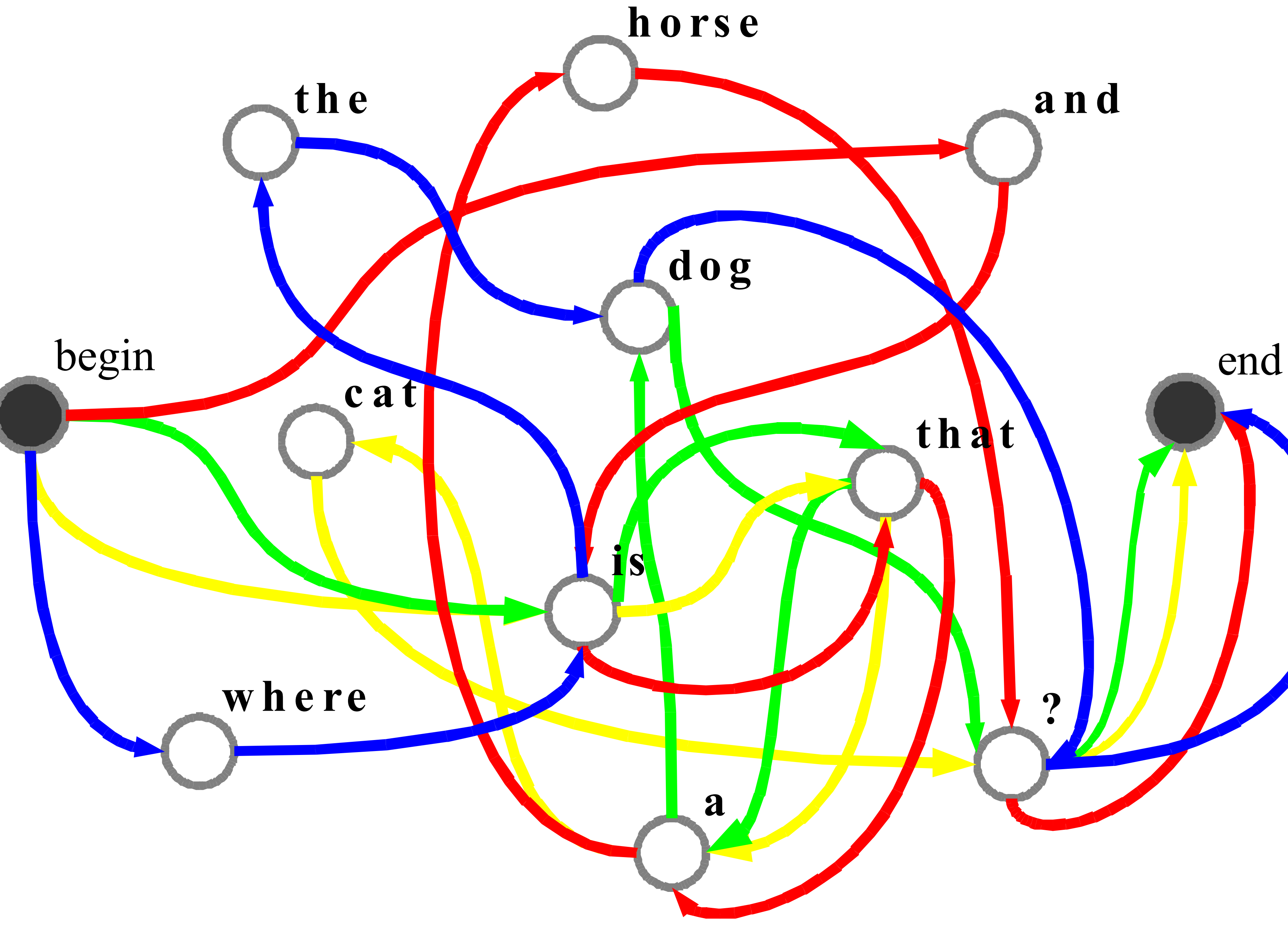
-computationally viable

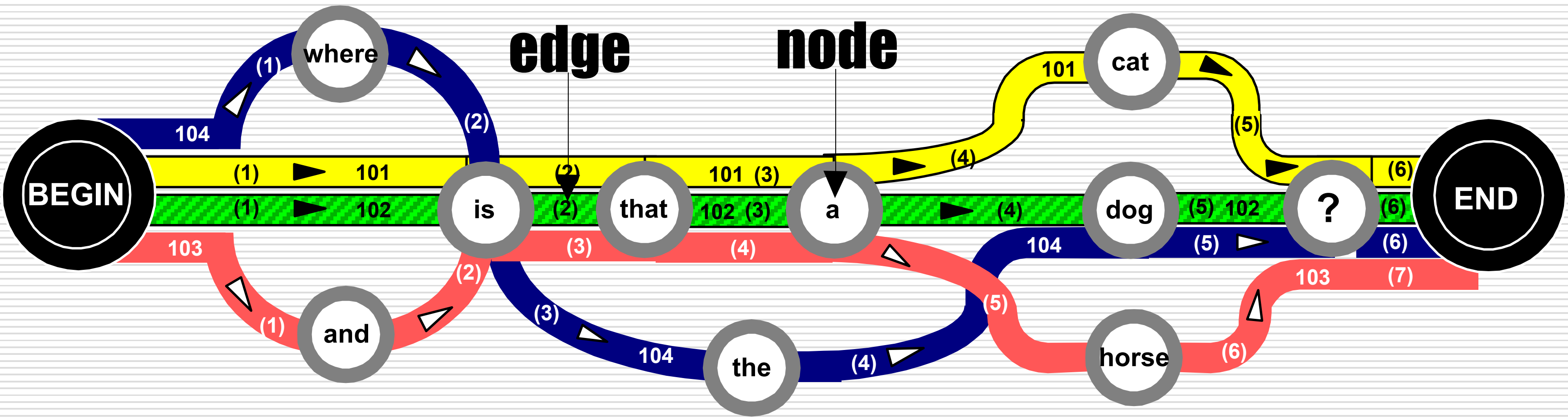
-theoretically sound

-empirically proven.

Inspiration: classical distributional approaches (Harris 1954, 1991), psycholinguistic data (Bates and Goodman 1999), grammar induction algorithms (Klein and Manning 2002), natural language processing (Barlow 2000).

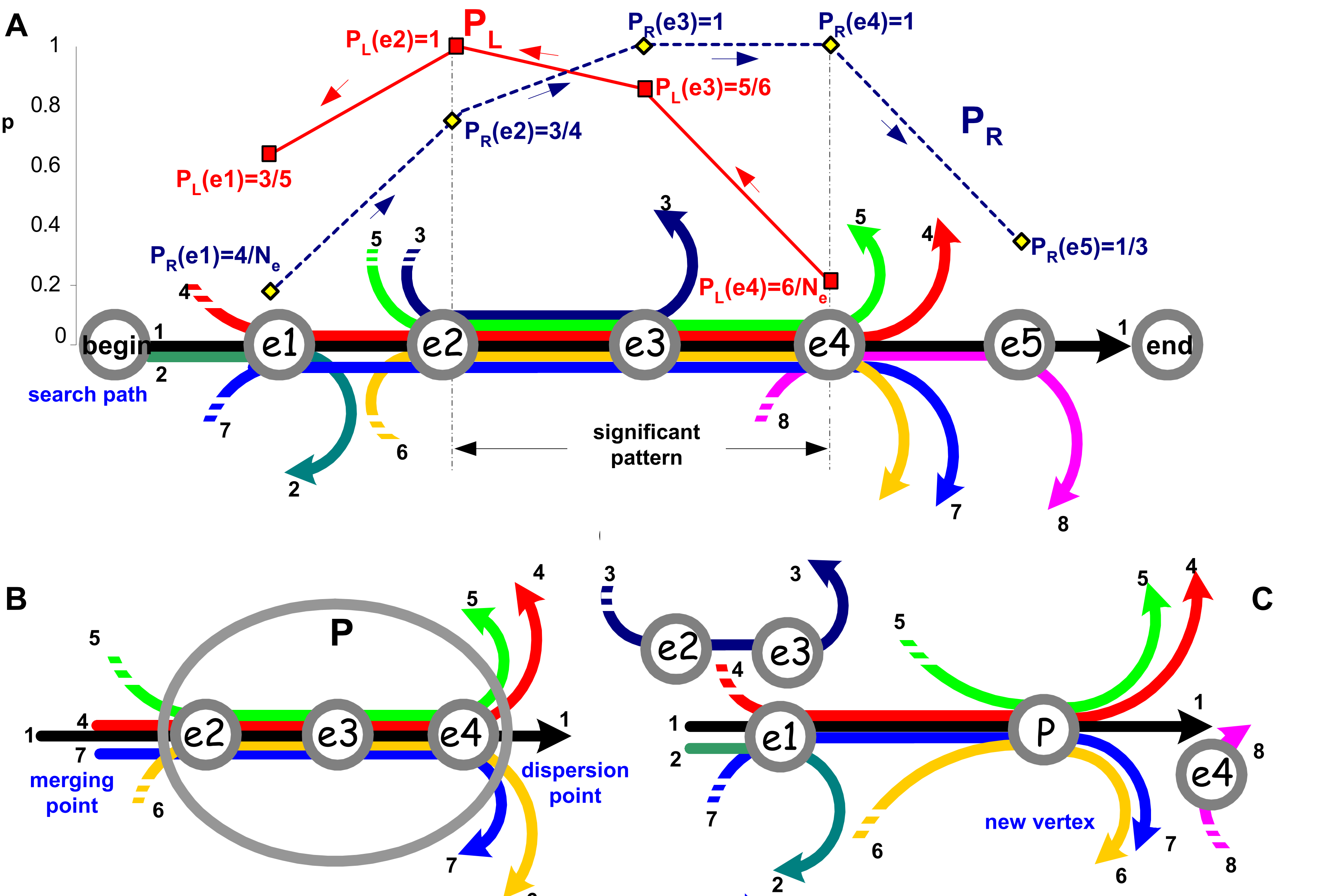
lexicon... and a graph induced by a corpus of strings



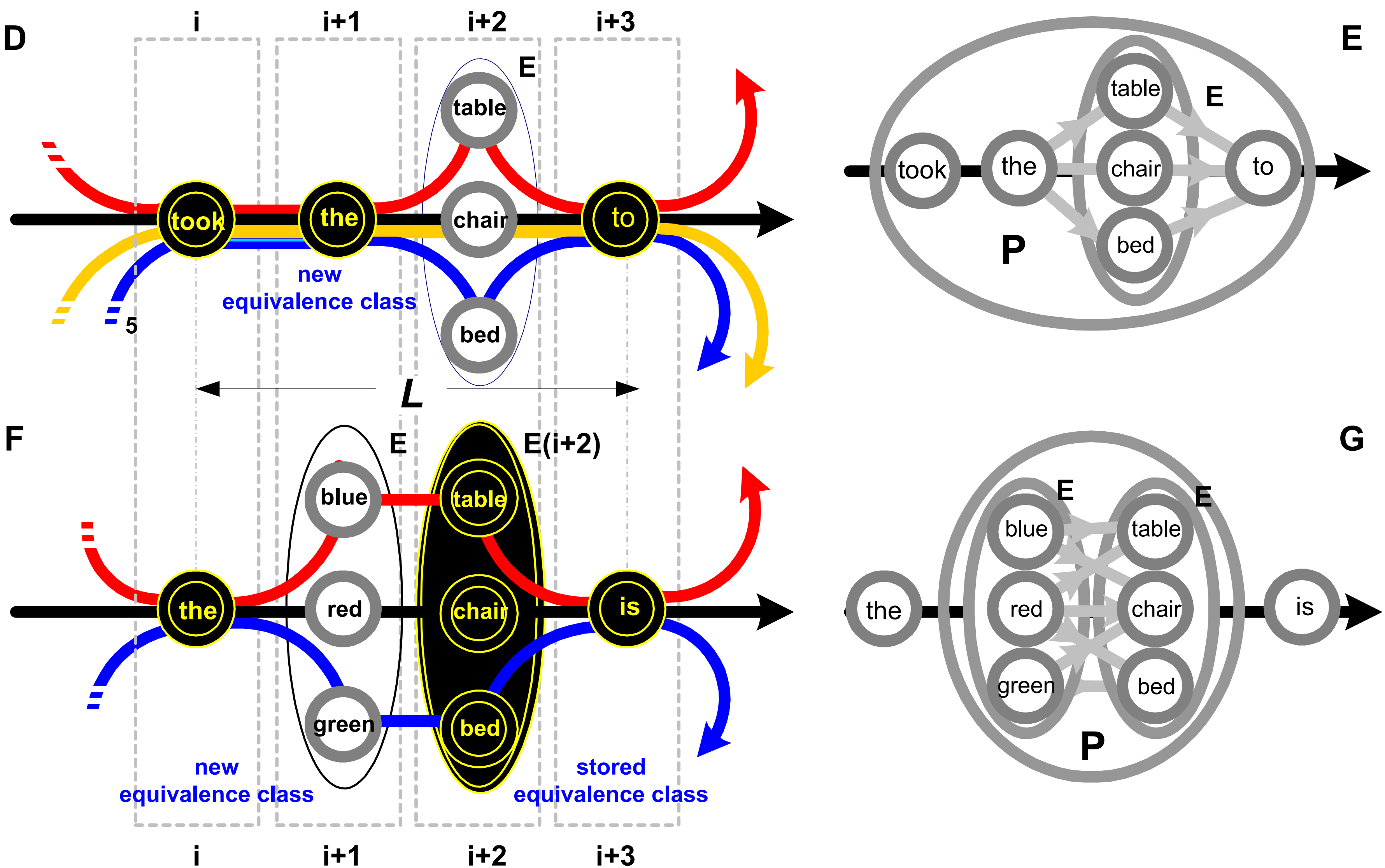


And is that a horse?

The MEX (motif extraction) procedure



Generalization



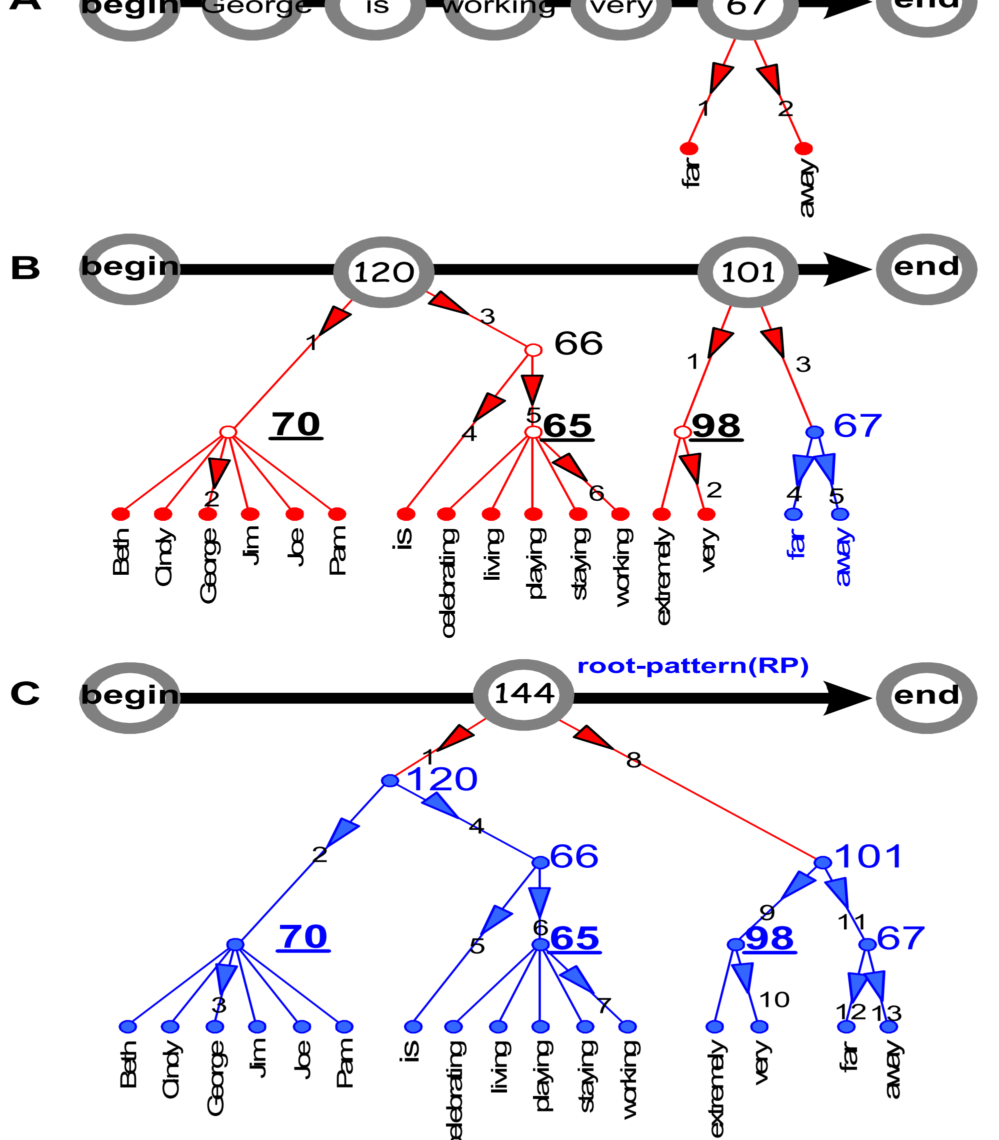
First pattern formation

Higher hierarchies: patterns (P) constructed of other Ps, equivalence classes (E) and terminals (T)

Trees to be read from top to bottom and from left to right

Final stage: root pattern

CFG: context free grammar



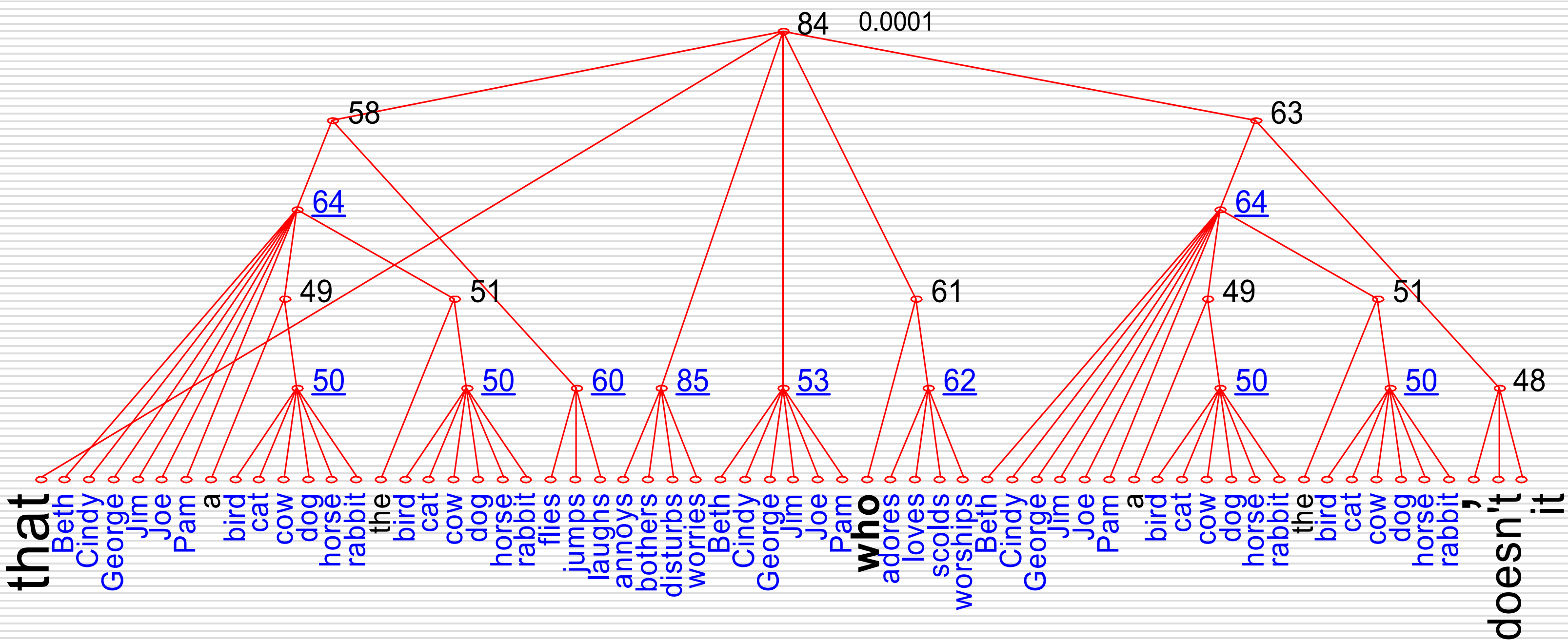
D

s ⟶ *begin* P144 *end*
P144 ⟶ P120 P101
P120 ⟶ E70 P66
E70 ⟶ *Beth* | *Cindy...*
P66 ⟶ *is* E65

P101 ⟶ E98 P67
E98 ⟶ *extremely* | *very*
P67 ⟶ *far away*

Two representations of a CFG

P84 → *that* P58 P63
P63 → E64 P48
E64 → *Beth* | *Cindy* | *George* | *Jim* | *Joe* | *Pam* | P49 | P5
P48 → *, doesn't it*
P51 → *the* E50
P49 → *a* E50
E50 → *bird* | *cat* | *cow* | *dog* | *horse* | *rabbit*
P61 → *who* E62
E62 → *adores* | *loves* | *scolds* | *worships*
E53 → *Beth* | *Cindy* | *George* | *Jim* | *Joe* | *Pam*
E85 → *annoys* | *bothers* | *disturbs* | *worries*
P58 → E60 E64
E60 → *flies* | *jumps* | *laughs*



that a bird flies bothers Jim who adores the cat, doesn't it

Example of context free grammar (first and last 15 out of 92 rules)

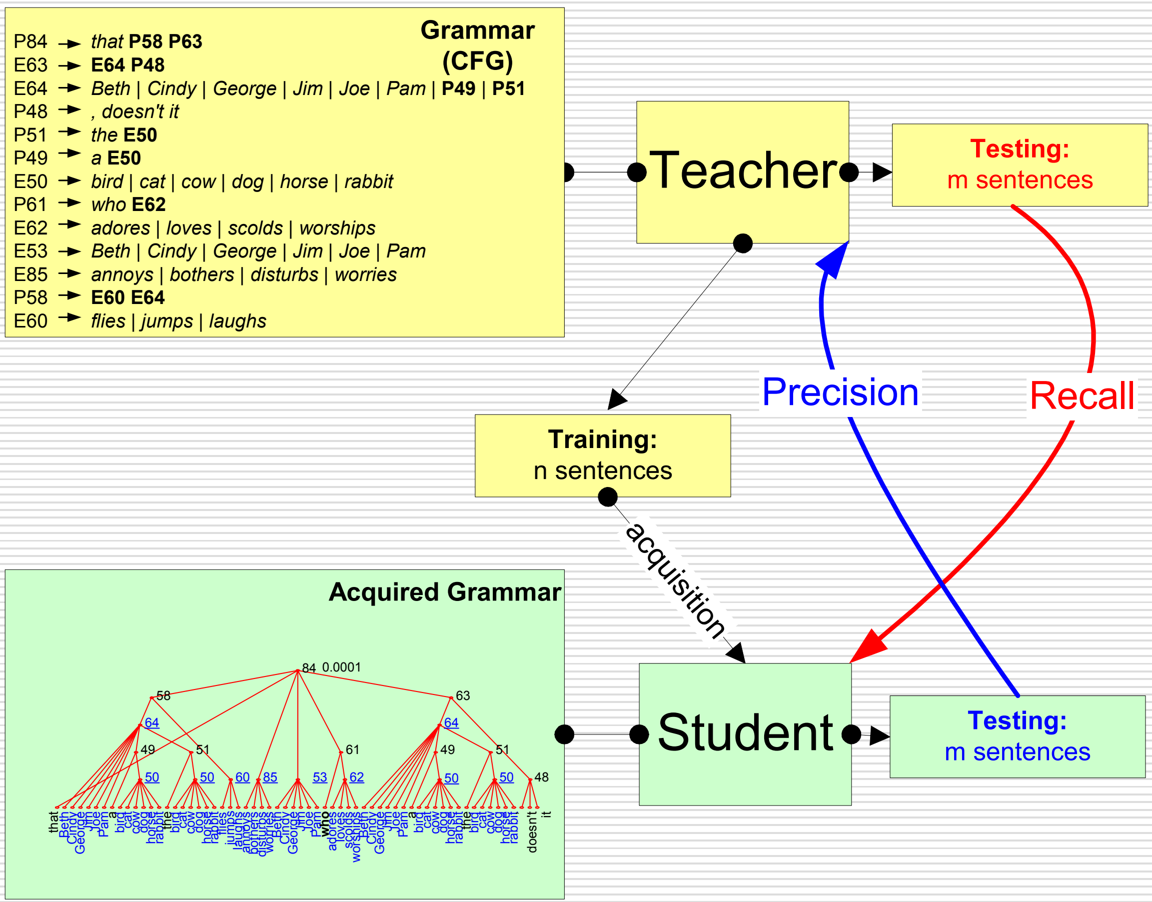
P53	(E54)	P130	(E131)
E54	{a,the}	E131	{P132,P133}
P55	(E56)	P132	(P73,P81,that,E131)
E56	{,}	P133	(P73,P81,that)
P57	(E58)	P134	(E135)
E58	{barks,meows}	E135	{P136}
P59	(E60)	P136	(P75,P81,that,E131)
E60	{flies,jumps,laughs}	P137	(E138)
P61	(E62)	E138	{P139}
E62	{that}	P139	(E101,that,E131)
P63	(E64)	P140	(E141)
E64	{annoys,bothers,disturbs,worries}	E141	{P142,P143,P144}
P65	(E66)	P142	(E138,E115,E128,E105)
E66	{eager,easy,tough}	P143	(E135,E95,E92)
P67	(E68)	P144	(P61,E115,E89,E95,P63,E115)

note loop

student learns from teacher

- Teacher generates a corpus of sentences
- Student distills syntax composed of significant patterns and equivalence classes
- Unseen teacher-generated patterns are checked by student (recall)
- Student-generated patterns are checked by teacher (significance)

student-teacher process



Recall=tp/(tp+fn)

Precision=tp/(tp+fp)

results

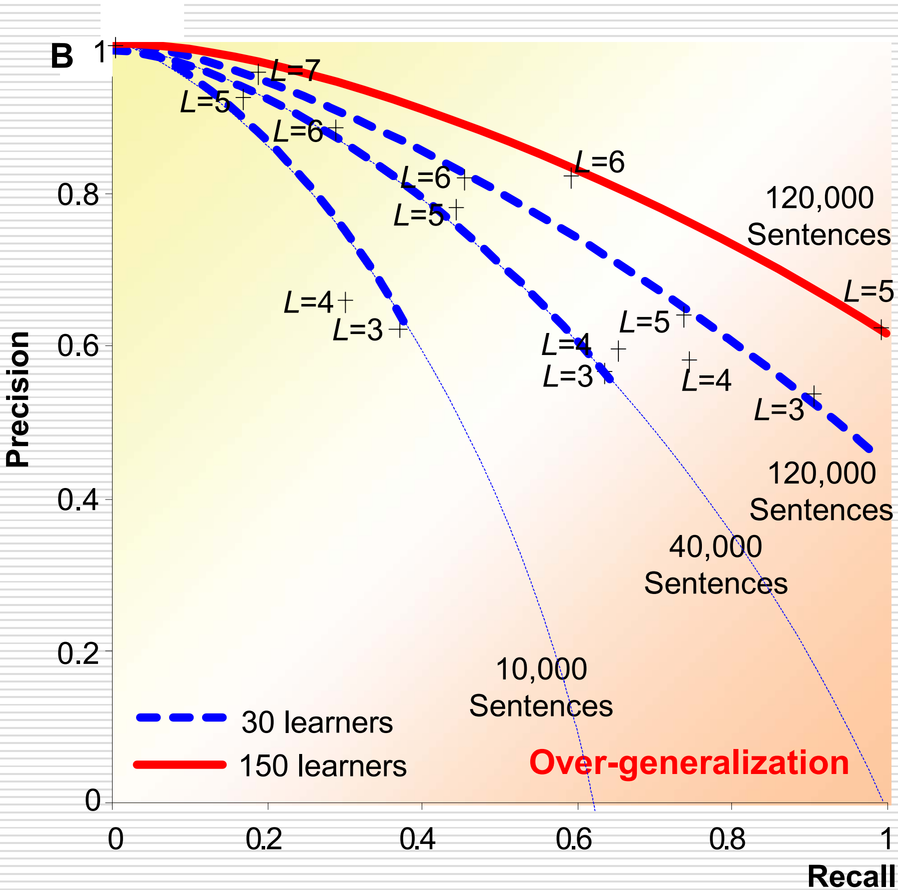
Corpus size	recall	error	precision	error
800	.85	.06	.72	.22
1600	.87	.06	.63	.09
3200	.84	.05	.61	.12
6400	.95	.01	.86	.08

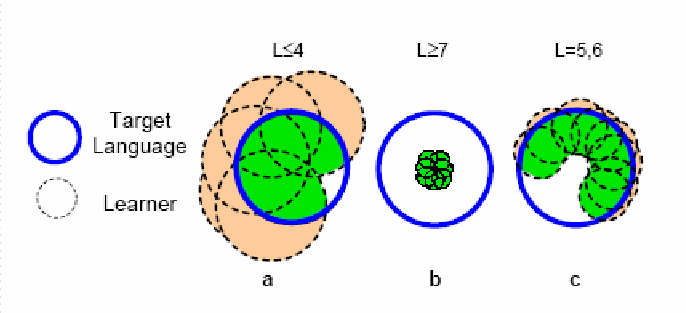
ATIS experiments

The ATIS-CFG is a hand-made CFG of 4592 rules, constructed to provide good recall (45%) of ATIS-NL, a corpus of natural language (13,000 sentences, 1300 words).

We train multiple ADIOS learners using ATIS-CFG as the teacher. Recursion is limited to depth 10. In testing performance, precision is defined by taking the mean across individual learners, while for recall acceptance by one learner suffices.

ADIOS learning from ATIS-CFG (4592 rules) using different numbers of learners, and different window length L





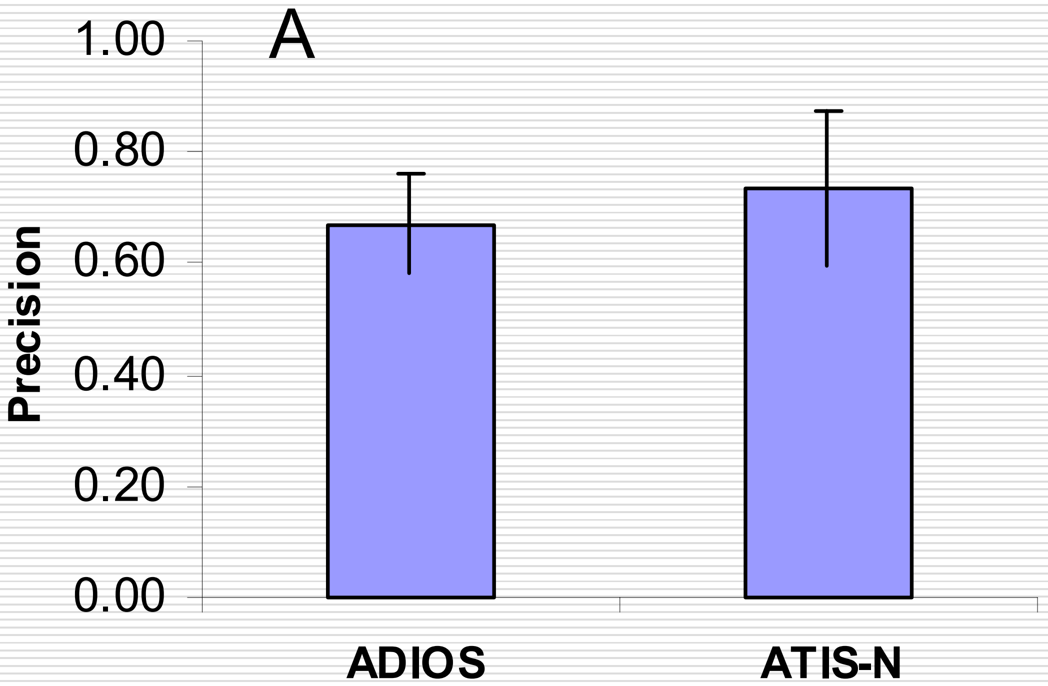
ATIS experiments

The ATIS natural language corpus contains 13,000 sentences.

Training ADIOS on it leads to recall of 40% (ATIS-CFG reaches recall of 45%). Nonetheless human judged precision is remarkable: 8 subjects judged the grammatical acceptability to be roughly the same as that of ATIS-NL!

All this while ATIS-CFG produces 99% of ungrammatical sentences!

Grammaticality



Language Model

•For any given l-word string

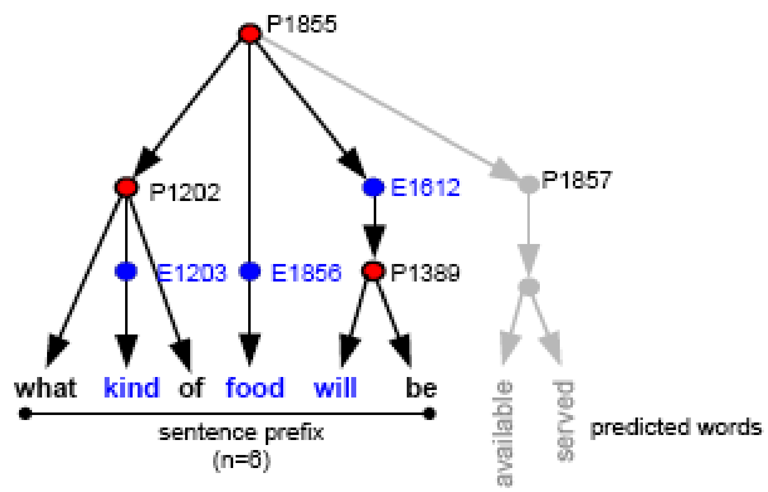
•For all prefixes of length k

•Find all parse-trees and assign probabilities for predicted words

•Assign larger weights to longer k

ATIS2 train:12.6K test:400

ATIS3 train:7K test: 1K



results

ATIS ver.	*method*	*# of parameters*	*perplexity*	*ref.*
2	ADIOS SSLM	under 5000	**11.5**	
2	Trigram Kneser-Ney backoff smooth.	1.E+05	14	[14]
2	PFA Inference (ALERGIA) + trigram	1.E+05	20	[15]
2	PFA Inference (ALERGIA)	1.E+05	42	[15]
3	ADIOS SSLM	under 5000	**13.5**	
3	SLM-wsj + trigram	1.E+05	15.8	[10]
3	NLPwin + trigram	1.E+05	15.9	[10]
3	SLM-atis + trigram	1.E+05	15.9	[10]
3	trigram	4.E+04	16.9	[10]
3	NLPwin	1.E+05	17.2	[10]
3	SLM-wsj	1.E+05	17.7	[10]
3	SLM-atis	1.E+05	17.8	[10]

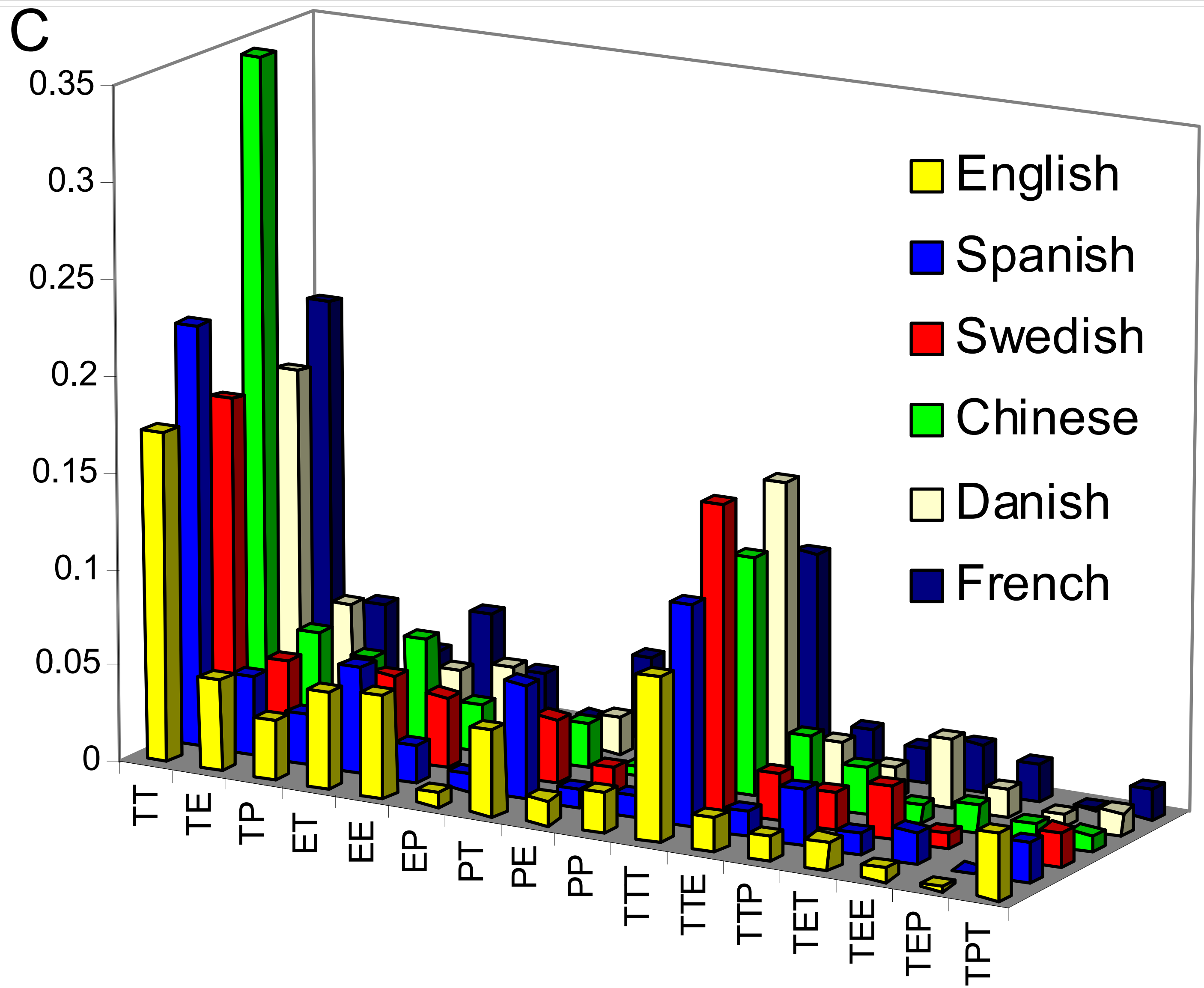
Table 1: The perplexity of the ADIOS SSLM, compared with some results from the literature [15, 14, 10]. Note that our SSLM uses for training *only* the data provided for that purpose in the ATIS corpora themselves. Although our model requires that only the three parameters of the ADIOS algorithm be specified in advance, we have stated the approximate overall number of patterns of all learners as the counterpart to the number of parameters in the other methods.

10: Chelba,2001. 14: Kneser-Ney 1995. 15: Kermovrant et al 2004.

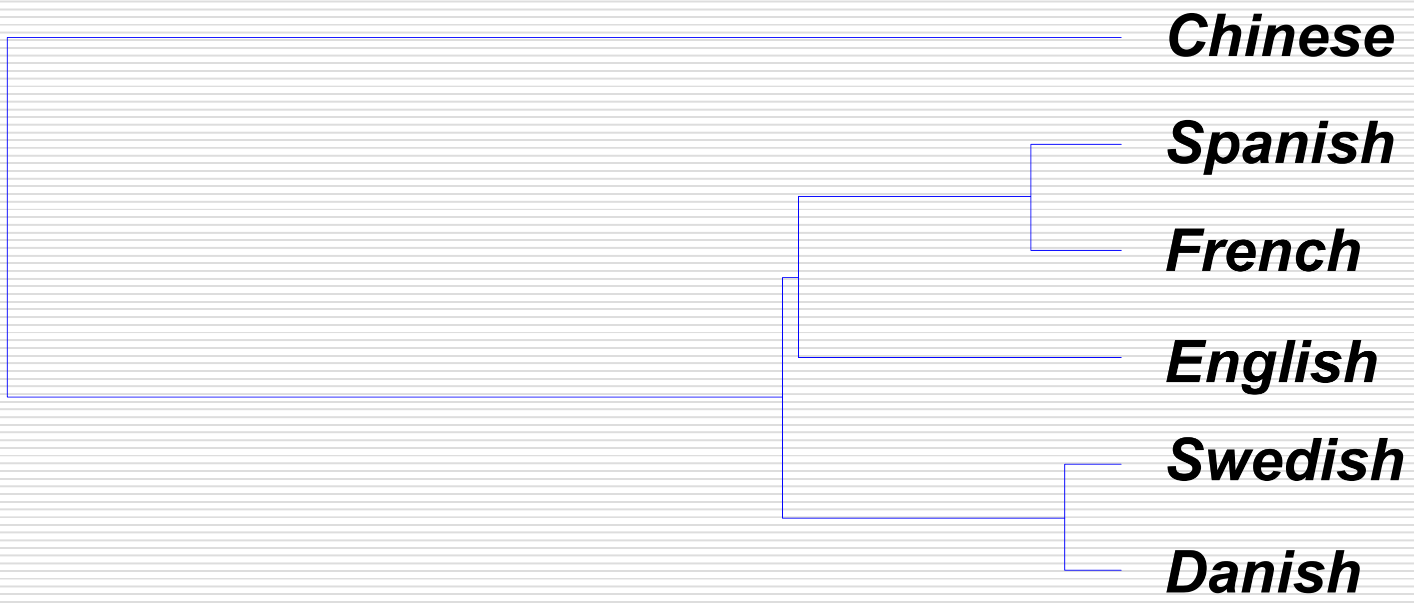
Meta-analysis of ADIOS results.

We define a pattern spectrum as the histogram of pattern types, whose bins are labeled by sequences such as (T,P) or (E,E,T), E standing for equivalence class, T for tree-terminal (original unit) and P for significant pattern.

We apply this analysis to the Parallel Bible, a text containing 31,000 verses in six different languages.

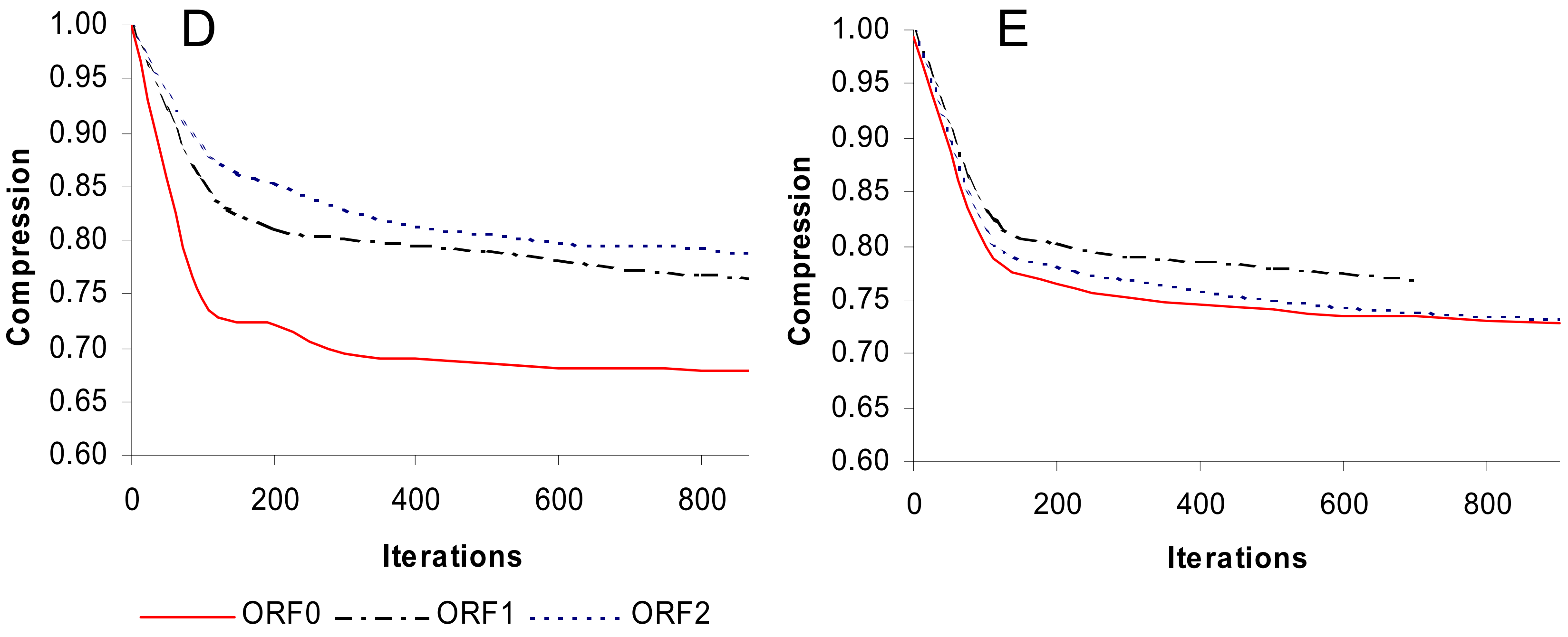

C
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
TT
TE
TP
ET
EE
EP
PT
PE
PP
TTT
TTE
TTP
TET
TEE
TEP
TPT
English
Spanish
Swedish
Chinese
Danish
French

Dendrogram of languages



The typological relations among six different natural languages, as judged according to pattern spectra. These are accepted relations according to linguists.

ADIOS analysis of 4777 genes in C. elegans, using as initial units the 64 codons (nucleotide triplets). D: first exon, E: first 500 bases



Compression is most favorable when correct Open Reading Frame is employed, in the coding case where it is meaningful.

Summary

- MEX is a motif-extraction method applicable to linguistic texts.
- Its application to proteins allowed for enzyme classification: from sequence to function!
- ADIOS is a grammar induction algorithm, employing MEX in a space of words, patterns and equivalence classes, and constructing a CFG representing syntax of the data.
- It was successfully applied to several corpora.
- It can serve as the basis for a language model.

Unsupervised learning of natural languages PNAS 102 (2005) 11629

http://horn.tau.ac.il http://adios.tau.ac.il