

m3it an3i3f .an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola aque5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2sh ar3ac ar5ial ar lnes 1nou lto 1tra 2de. 2dly 21m 21out 2p1t 2p2ec ool3ic bor ou3re bu4 co3e co3: ct5ang cuc 4rin c4t de3ra .de dre4 dril eal3ou ea ea2v ea4ge ede4s edĭ en3g en3 ep4sh ep5a

#### What is this lecture about

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. a :d5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1 al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana ande4s .ang4 ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1d1 an1g1 an2sa an2sp an2tr an3age an3arc an3dis an3io an3i

Generation of patterns

How T<sub>F</sub>X hyphenates words

Using PATGEN

Why PATGEN doesn't suit us today

- PATLIB and OPATGEN, system architecture
- Pattern recognition in typesetting
- Future applications

a3ta fa3the fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi: i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c fril4 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les

e4dler e4fic e4fuse, e4f3ere e4go, e4gos e4jud e4la, e4lac e4law e4led e4ller e4l3ing e4lbic, e4mac e4mag e4met e4mis e4mul e4n odnos edoid edot edpli edprec edpred edprob edput edq3ui3s edriva edsage. edsages\_edsert. edserts edserva\_edvin\_edwag e5and e5a ex e pits épur Nic es pst est Backs Nexteng es verb es voc es vu es vec fab 3r fain 4 fambis l'as fab de la fab upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola ague5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2<mark>s</mark>h How T<sub>E</sub>X hyphenates words Before 2<sup>nd</sup> pass of paragraph breaking, all words TEX finds all *patterns* of the word, all hyphens nes 1nou 2de. 2dly Pattern is a subword, with hyphenation information between symbols 2p1t 2p2ed n2at hen5at Hyphenation information—numbers odd—allow breaking—covering even—disallow breaking—inhibiting Patterns compete—higher value wins ep4sh ep5

eddler e4fic e4fuse. e4f3ere e4go. e4gos e4jud e4la. e4lac e4law e4led e4ller e4l3ing e4lbic. e4mac e4mag e4met e4mis e4mul e4n e4nos e4oi4 e4ot e4pli e4prec e4pred e4prob e4put e4q3ui3s e4riva e4sage. e4sages\_e4sert. e4serts e4serva\_e4vin\_e4wag e5and e5a

i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c fril4 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fle f4fly f4l2 f4to f5fin. f5les

epits epur Nic espect esi Beslen Nilim esloc eslud esman esmiss esnea esnea esnia esiil esnia 15 esout esow el A sereveg de Nextegitio Backs Nexteng esverb esvoc esvu eswee fab3r fain4 fQuitfam5i2/15 faults fa3b; E fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. ac ad5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1: al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana

ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1dl an1g1 an2sa an2sp an2tr an3age an3arc an3dis an3io an3io an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apoc

upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola aque5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2sh ar3ac ar5ial ar lnes 1nou co3e co3: ea2v ea4g ede4s edi en3g en3 ep4sh ep5a

## **Example of application of patterns**

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. a :d5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1 al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana ande4s .ang4 ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1d1 an1gI an2sa an2sp an2tr an3age an3arc an3dis an3io an3 m3it an3i3f .an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor

- hyphenation 1n a 1t i o n2a t 2i o h e2n .h y3p h hena4 h e n5a t
- h y-p h e n-a t i o n

.h0y3p0h0e2n5a4t2i0o0n.

- hyphenation point occurrence
  - intelligence (magic?)

e4dler e4fic e4fuse, e4f3ere e4go, e4gos e4jud e41a, e41ac e41aw e41ed e411er e413ing e41bic, e4mac e4mag e4met e4mis e4mul e4n e4nos e4oi4 e4ot e4pli e4prec e4pred e4prob e4put e4q3ui3s e4riva e4sage. e4sages e4sert. e4serts e4serva e4vin e4wag e5and e5a escite esex espits pur Nic appst esi Backs Nexteng esverb esvoc esvu eswee fab3r fain4 fQuitfam5i3/15 esout esow esserb esvec esvu eswee fab3r fain4 fQuitfam5i3/15 h faults fa3b a3ta fa3the fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi: i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c fril4 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les

#### Patterns hold context of

Pattern recognition—a kind of

.an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola ague5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2<mark>s</mark>h Pattern generating process We want *minimal* set of patterns Generating completely covering minimal patterns is NP-complete nes 1nou to 1tra 2de. 2dly Good iterative methods for "near optimal" solution 2p1t 2p2ed One way to do it: dic-tio-nary with marked hy-phen points we repeat going through the dictionary making levels *odd*—covering even—inhibiting ep4sh ep5 e4dler e4fic e4fuse. e4f3ere e4go. e4gos e4jud e4la. e4lac e4law e4led e4ller e4l3ing e4lbic. e4mac e4mag e4met e4mis e4mul epits Pyronic et pet e51 Backs Nexteng e5verb e5voc e5vu e5wee fab3r fain4 fQuitfam5i.4/15 e6v

abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. :d5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1 al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana

fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c fril4 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les

at5ac at5a avi4er avi	<ul> <li>candidate choosing rule—we take subwords (k-character subword)</li> </ul>	u5sib ave
a3nen a3ne a4m5ato a4 a5nur a5ra	$\star$ and count the number of their good and bad work	i4ly. a4ma inimi a5ni parbi5 ban
bas4e bas be5gu be5r blath5 ble 21er z4il	* hen(35,4)a	estw best t bi5ou h co5ol zte4 leff lexy
fa 1fi 11 nes 1nou to 1tra 1 2de. 2dlv	<ul> <li>not all the candidates are good</li> </ul>	1na 1nen tiv 1tiz 2c5ah 2da 1 2gue 2gf
2h1n 2ch': 2lm 2lout 2p1t 2p2ec bol3ic bor bu3re bu4g an3iz car ci4a ccor bi5c ck1 (	<ul> <li>         * pattern choosing rule—linear function over the number of good and bad work of the pattern compared to a threshold:         good_count * good_wt − bad_count * bad_wt ≥ threshold     </li> </ul>	to 21d 211 id 2oph 2r; bne5g bc it4ti bu3cam4pe c a5den cat .3ph ci41a cow5a coz
co3e co3: ct5ang cuc c4rin c4t: lemor5 der	<ul> <li>we put good candidates into pattern set with current level number</li> </ul>	.na c4one deli4e de iu .de3o d
lis1 di1re lre4 dri14 l1uca d1u1	∗ hen5a	n drea5r l1j d1m d1 k2 ead1 e
eal30u eal ea2v ea4ge ede4s edi ef5i5nite ello4 el21 emu3n em1: en3g en3i	<ul> <li>they still make errors, next level will correct errors of patterns selected so far</li> </ul>	sifie ec5i efil4 efc d elaxa4 moni5o en 13em en3et
py4sh ep5a er1i er1ou es5cu es5a et5rif et4 ew3ing .ev e1wa e2col	<ul> <li>exception list at the end of the whole process, to correct remaining errors</li> </ul>	; er1a er?; ;4w es5car ;t5itiv et et1ic ewi .s4a e1vi e3imb e3i e3tra e3t
Sup e3wh 4dler e4fi 4nos e4oi4 5cite e5ex 5roc e5ski 6a3ta fa3th i3a fi3cer rol5 fusi4	ic e4fuse. e4f3ere e4go. e4gos e4jud e4la. e4lac e4law e4led e4ller e4l3ing e4lbic. e4mac e4mag e4met e4mis 4 e4ot e4pli e4prec e4prec e4prob e4put e4q3ui3s e4riva e4sage. e4sages e4sert. e4serts e4serva e4vin e4wag x epit5 epur nic e5pst e5ip e5ten flim e5loc e5lud e5man e5miss e5nea e5nea e5nea e5nig e5il e5nig e5pst e5o e5vent e5voc e5vu e5wee fab3r fain4 fQuitfam5i5/a15h fahe fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n r fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c f fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5f	e5and e5at out e5ow e8 oult5 fa3bl fis4ti fi2 cril4 fri2

wbe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad41e ad4su ad5er. acad5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah41 .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1: al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ans ande4s .ang4 ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1dl an1gl an2sa an2sp an2tr an3age an3arc an3dis an3ic an3 an3it an3i3f .an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est am5ot apar4 apoc5 apo apos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola ague5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2sh ar3ac

r5n ate5

ar3al ar3∈

ar5ial ar

How good the generation process is Input data  $\approx$  MB patterns  $\approx$  tens of KB lnes 1nou to 1tra 2de. 2dly covering > 98 %21m 21out

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. a :d5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1 al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana ande4s .ang4 ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1d1 an1g1 an2sa an2sp an2tr an3age an3arc an3dis an3io an3i m3it an3i3f .an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola aque5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2sh ar3ac

with error < 0.1 %

2p1t 2p2ed

co3e co3:

de3ra .de

dre4 dril

ea2v ea4ge

en3g en3 ep4sh ep5a

Results depend on parameter setting

usual number of levels  $\leq 5$ 

- Nobody knows how to set the parameters

e4dler e4fic e4fuse, e4f3ere e4go, e4gos e4jud e4la, e4lac e4law e4led e4ller e4l3ing e4lbic, e4mac e4mag e4met e4mis e4mul e4n anos e4014 e40t e4pli e4prec e4pred e4prob e4put e4q3ui3s e4riva e4sage. e4sages e4sert. e4serts e4serva e4vin e4wag e5and e5an escite esex espits pur Nic espet esi Backs Nexteng esperb espoc esua esman esmiss esnea esnee esniques il esnic 15 espot esow esserb espec espoc espot espec fab3r fain4 fQuitfam5i.6/15 h faults fa3b

a3ta fa3the fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi: i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c fril4 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les

**PATGEN** Frank Liang, 1982, 7-bit ASCII

lnes 1nou

co3e co3 4rin c4t

en3g en3 ep4sh ep5a

### **Generating with PATGEN**

ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1dl an1gl an2sa an2sp an2tr an3age an3arc an3dis an3io an3io an3io an5io an5 .an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. a :d5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1 al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana

upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola aque5 aran4¢ araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2b ar2sh

- Yannis Haralambous, Karl Berry,...)

later modified for accents, 8-bit ASCII,... (Peter Breitenlohner,

Patterns for tens of languages exist

from a dictionary

first levels by hand

e4dler e4fic e4fuse, e4f3ere e4go, e4gos e4jud e4la, e4lac e4law e4led e4ller e4l3ing e4lbic, e4mac e4mag e4met e4mis e4mul e4n odnos edoid edot edpli edprec edpred edprob edput edq3ui3s edriva edsage. edsages\_edsert. edserts edserva\_edvin\_edwag e5and e5a epits pyr Nic et est est a slen Nexteng esterb este a3ta fa3the fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi

i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c fril4 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les lnes 1nou lto 1tra 2de. 2dly 2h1n 2ch 2p1t 2p2ec ool3ic bor co3e co3: 4rin c4t en3g en3 ep4sh ep5a

# Why PATGEN does not suit us today

ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1dl an1g1 an2sa an2sp an2tr an3age an3arc an3dis an3io an3io an3io an5io an5 .an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. a :d5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1 al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana

upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola aque5 aran4¢ araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2b ar2sh

Monolithic structured code, difficult to maintain

Only ASCII, approx. 240 symbols

 $\Omega$  handles UNICODE

- Static data structures
- Difficult to use for purposes other than hyphenation

e4dler e4fic e4fuse, e4f3ere e4go, e4gos e4jud e4la, e4lac e4law e4led e4ller e4l3ing e4lbic, e4mac e4mag e4met e4mis e4mul e4n odnos edoid edot edpli edprec edpred edprob edput edq3ui3s edriva edsage. edsages\_edsert. edserts edserva\_edvin\_edwag e5and e5a 

a3ta fa3the fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi: i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer~fos5 fo2r fo5rat fra4t fres5c fri14 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola aque5 aran4¢ araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2b ar2sh PATLIB and OPATGEN PATtern manipulating LIBrary—from scratch O  $(\Omega?)$  PATtern GENerator—handles UTF-8 UNICODE encoding lnes 1nou lto 1tra 2de. 2dly Generalisation of PATGEN 21m 21out 2p1t 2p2ed ool3ic bor Implementation: C++ in CWEB co3e co3 http://www.fi.muni.cz/~xantos/patlib c4rin c4t de3ra .de dis1 di1re **GPL** dre4 dril ea2v ea4ge en3g en3 ep4sh ep5a e4dler e4fic e4fuse, e4f3ere e4go, e4gos e4jud e4la, e4lac e4law e4led e4ller e4l3ing e4lbic, e4mac e4mag e4met e4mis e4mul e4n odnos edoid edot edpli edprec edpred edprob edput edq3ui3s edriva edsage. edsages\_edsert. edserts edserva\_edvin\_edwag e5and e5a ex e pits épur Nic es pst est Backs Nexteng es verb es voc es vu es vec fab 3r fain 4 fQuitfam 5 i 9 a fain faults fasb a3ta fa3the fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi; i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c fril4 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. ac ad5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1: al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana

ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1dl an1gl an2sa an2sp an2tr an3age an3arc an3dis an3io an3io an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor

**PATLIB** Finite language store (pattern manipulator) implemented using packed trie (see the Proceedings) insert pattern

#### Generator

- easy to change
- OPATGEN is in fact "only" the word list input/output interface

rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les

Bits Pg Nic st Pst e51 Besten Niemte51oc e51ud e5man e5miss e5nea e5nee e5

PATLIB architecture

:d5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1 .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1dl an1gI an2sa an2sp an2tr an3age an3arc an3dis an3io .an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor

pos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola aque5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2sh

ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er.

- - delete pattern
  - ...low-level operations
- - handles creating patterns using the same strategy as PATGEN

fic e4fuse. e4f3ere e4go. e4gos e4jud e4la. e4lac e4law e4led e4l1er e4l3ing e4lbic. e4mac e4mag e4met e4mis e4mu

e4ot e4pli e4prec e4pred e4prob e4put e4q3ui3s e4riva e4sage. e4sages e4sert. e4serts e4serva e4vin e4wag e5and i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer~fos5 fo2r fo5rat fra4t fres5c fril4 fri2 ad5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1: al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana ande4s .ang4 ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1d1 an1g1 an2sa an2sp an2tr an3age an3arc an3dis an3io an3i .an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola aque5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2sh ar3ac ar5ial ar **Pros** and cons What do we get implementation using logical operations lnes 1nou lto 1tra 2de. 2dly any strategy of generating is easier to implement 2p1t 2p2ed generality and flexibility co3e co3 ct5ang cuc user manual :-) c4rin c4t de3ra .de What do we pay for it dre4 dril ea2v ea4ge performance loss ede4s edi en3g en3 ep4sh ep5a e4dler e4fic e4fuse, e4f3ere e4go, e4gos e4jud e4la, e4lac e4law e4led e4ller e4l3ing e4lbic, e4mac e4mag e4met e4mis e4mul e4n anos e4014 e40t e4pli e4prec e4pred e4prob e4put e4q3ui3s e4riva e4sage. e4sages e4sert. e4serts e4serva e4vin e4wag e5and e5an ex e Bits Pur Nic et Pst e51 Backs Nexteng e5verb e5voc e5vu e5wee fab3r fain4Quit fam511/15h a3ta fa3the fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi: i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer~fos5 fo2r fo5rat fra4t fres5c fril4 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. a

Hyphenating point → point of interest
 Eogu best base best base best laths bit laths l

May be implemented as Pattern Translation Process (PTP)

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad41e ad4su ad5er. ac ad5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai51y ak1en ali4e al1:

Generalisation, abstraction

pos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola ague5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2sh

.al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 anidl anigl an2sa an2sp an2tr an3age an3arc an3dis an3io an3i an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor

• Some can be done as  $\Omega TP$  in Omega

"word length"

ep4sh ep5

e31mb e34ra e4fuse. e4f3ere e4go. e4gos e4jud e4la. e4lac e4law e4led e4ller e4l3ing e4lbic. e4mac e4mag e4met e4mis e4mul e4oi4 e4ot e4pli e4prec e4pred e4prob e4put e4q3ui3s e4riva e4sage. e4sages e4sert. e4serts e4serva e4vin e4wag e5and ee5ex e5ji5 e5jur Nic e5pst e5i Be5len Nilm e5loc e5lud e5man e5miss e5nea e5nea e5pec e5nil e512/15 ebout e5ow e5skin BdevFsgdNextPsgitioBacksNexteng e5verb e5voc e5vu e5wee fab3r fain4Quite fam512/15 h fault5 fa

i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c fril4 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les

- we want to prefer breaking on word boundaries

- we want to prefer breaking on word boundaries

- we want to prefer breaking on word boundaries

- needed for German (Wort-silben-trennung)

- Context dependent ligatures

- ligatures over compound word boundaries are wrong

- ligatures over compound word boundaries are wrong

- English: shelfful vs. shelfful

- Czech: šéflékař vs. šéflékař (doctor in chief)

Adler e4fic e4fuse. e4f3ere e4go. e4gos e4jud e4la. e4lac e4law e4led e4l1er e4l3ing e4lbic. e4mac e4mag e4met e4mis e4mul

e4oi4 e4ot e4pli e4prec e4pred e4prob e4put e4q3ui3s e4riva e4sage. e4sages e4sert. e4serts e4servă e4vin e4wag e5and e5ar e e5ex epit5 epur Nic e5pst e5iB e5len Nilm e5loc e5lud e5man e5miss e5nea e5nea e5nei e5nil e513/15 e5out e5ow e e5skin B4EVF5gdNextPegitioBack5Nexteng e5verb e5voc e5vu e5wee fab3r fain4Quite fam513/15h fault5 fa3b

i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c fril4 fri2 Fro15 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f412 f4to f5fin. f5les

fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. ac ad5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1: al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana

**Examples of usage** 

upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola ague5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2sh

Hyphenation of compound words

Fraktur long s versus short s

morphology dependent

ep4sh ep5

ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1dl an1g1 an2sa an2sp an2tr an3age an3arc an3dis an3io an3io an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor

ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1dl an1gl an2sa an2sp an2tr an3age an3arc an3dis an3io an3io an3io an5io an5 .an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola ague5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2sh End of sentence different space width  $\@$  and  $\U$  in LATEX (sorry, the slides are in ConTEXt) lnes 1nou Thai segmentation 21m 21out 2p1t 2p2ed word/sentence boundaries are not present in the input Thai transcription, needed for line-breaking co3e co3: Arabic letter hamza five hamza variants, depending on context ea2v ea4g Adding Greek accents Adding Czech accents to 7-bit ASCII texts en3g en3 ep4sh ep5a e-mail,... e4dler e4fic e4fuse. e4f3ere e4go. e4gos e4jud e4la. e4lac e4law e4led e4ller e4l3ing e4lbic. e4mac e4mag e4met e4mis e4mul Anos e4oi4 e4ot e4pli e4prec e4pred e4prob e4put e4q3ui3s e4riva e4sage. e4sages e4sert. e4serts e4serva e4vin e4wag e5and e5an epits Province Province to Backs Nexteng e5verb e5woc e5vu e5wee fab3r fain4 Quit fam514/15h fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer fos5 fo2r fo5rat fra4t fres5c fril4 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. ac ad5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1: al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana



Performance loss 10–15 times to PATGEN Caused by lnes 1nou lto 1tra 2de. 2dly breaking the program into logical layers using dynamic memory co3e co3 alphabet symbol is an object, not mere number 4rin c4t not (yet, if ever) optimised ea2v ea4ge en3g en3 ep4sh ep5a e4dler e4fic e4fuse, e4f3ere e4go, e4gos e4jud e4la, e4lac e4law e4led e4ller e4l3ing e4lbic, e4mac e4mag e4met e4mis e4mul e4n e4nos e4oi4 e4ot e4pli e4prec e4pred e4prob e4put e4q3ui3s e4riva e4sage. e4sages e4sert. e4serts e4serva e4vin e4wag e5and e5an epits pur Nic epit esi Back Nexteng esverb esvoc esvu eswee fab3r fain4 a3ta fa3the fa4ce feas4 feath3 fend5e fen2d fer1 .fes3 fev4 fe3li fe4b fe4mo fic4i fight5 fill5in fil5i fin2d5 fin4n fis4ti fi i3a fi3cer fi3cu fi5del flin4 flo3re fon4de fon4t fore5t fort5a for4i for5ay .for5mer~fos5 fo2r fo5rat fra4t fres5c fri14 fri2 rol5 fusi4 fus4s fu3ri fu4min fu5el fu5ne f1in3g f2fy f2f5is f2ly5 f2ty f3ican f3icen f4fes f4fie f4fly f4l2 f4to f5fin. f5les

abe2 abi5a ab3ul ab5erd ab5it5ab ac1in ac3ul ac4um ac5ard ac5aro ac5rob adi4er ad3ica ad3ow .ad4der ad4din ad4le ad4su ad5er. ac ad5um aeri4e ae4r aff4 .af1t aga4n age4o ag1i ag1n ag3oni ag5ell ag5ul ah4l .ach4 ain5in ain5o ait5en ai2 ai5ly ak1en ali4e al1: al3ad al3end .al3t al4ia. al5ab al5lev ama5ra ami4no amor5i amp5en am1in am3ag am3ic am5ab am5asc .am5at am5era am5if am5ily ana

upos3t aps5es ap3in ap3ita ap5at ap5ero ap5illar ap5ola ague5 aran4g araw4 ara3p arbal4 arre4 ar1i ar2iz ar2mi ar2p ar2sh

ang5ie .ani5m anoth5 ano4 ans3po antal4 .anti5s .ant4 an1dl an1gl an2sa an2sp an2tr an3age an3arc an3dis an3io an3io an3te an3ua an3ul an4dow an4ime an4kli an4sco an4sn an4st an4sur an4tie an4tw .an5c an5est. an5ot apar4 apoc5 apor