Syntactic Change and Typology

Ted Briscoe

Computer Laboratory
Natural Language and Information Processing Group
University of Cambridge

ENS, Paris
Mar 2014

Statistical Universals/Commonalities

- Typology: 6K attested lgs (1K in Papua New Guinea!) – control for geography and history
- Word Order: SVO, SOV $>$ VSO $>$ VOS $>$ OVS, OSV
- Correlations: OV $\leadsto$ Rel+N $\wedge$ Case Marking $\wedge$ Postpositions
- Kim ga kiss Sandy wa Robin ga kiss
  Sandy who kissed Kim kissed Robin
- Irregularity / Frequency: irregular (less-productive) forms are more frequent *go+ed / went, travel+ed / *travd
  A+N & N+A (French)

Statistical Universals/Commonalities

- Typology: 6K attested lgs (1K in Papua New Guinea!) – control for geography and history
- Word Order: SVO, SOV $>$ VSO $>$ VOS $>$ OVS, OSV
- Correlations: OV $\rightsquigarrow$ Rel+N $\wedge$ Case Marking $\wedge$ Postpositions
- Kim ga kiss Sandy wa Robin ga kiss
  Sandy who kissed Kim kissed Robin
- Irregularity / Frequency: irregular (less-productive) forms are more frequent *go+ed / went, travel+ed / *travd
  A+N & N+A (French)

Typology – Method

1. About 6k attested languages (1k in New Guinea!)
2. Half the world's population speaks (natively) languages which have developed from Proto-Indo-European (140 total; but English, Spanish,... more popular than Hittite, Dutch,...)
3. Typologists study samples balanced for geographical and historical relationships (lg families, lg contact)
4. Massive skewing in parametric combinations: statistical and implicational universals

Subj-Vb-Obj order:
SVO 42%, SOV 45%, VSO 9%, VOS 3%, OVS 1%, OSV

VO $\rightarrow$ PrepPos $\wedge$ Aux-Vb $\wedge$ N-RelCl

OV $\rightarrow$ PostPos $\wedge$ Vb-Aux $\wedge$ Case-marking

RelCl-N $\rightarrow$ PostPos

(Universal) Shift-Reduce Parsing Procedure

1. The Reduce Step: if the top 2 cells of the stack are occupied, then try
   a) Application, if match, then apply and goto 1), else b),
   b) Composition if match then apply and goto 1), else c),
   c) Permutation, if match & new, then apply and goto a), else goto 2)
2. The Shift Step: if the first cell of the Input Buffer is occupied, then pop it and move it onto the Stack together with its associated lexical syntactic category and goto 1),
   else goto 3)
3. The Halt Step: if only the top cell of the Stack is occupied by a constituent of category S,
   then return Success,
   else Fail

1-1 Bounded Context Shift-Reduce Parse

Stack (PDS)	Input Buffer	Operation
	Kim loves Sandy	
Kim:NP:kim′	loves Sandy	Shift
loves:(S\NP)/NP:λ y,x love′(x y) Kim:NP:kim′	Sandy	Shift
Kim loves:S/NP:λ y love′(kim′ y)	Sandy	Reduce (P,BA)
Sandy:NP:sandy′ Kim loves:S/NP:λ y love′(kim′ y)		Shift
Kim loves Sandy:S:love′(kim′ sandy′)		Reduce (FA)

Learning via Parse Failure

- Parse with current parameter settings (P-settings)
- If Learning LAgt & Parse Failure, then Update P-settings
- Assume fm_i, then valid category assignment (VCA) to i
- Kim kisses Sandy : Kiss′(kim′,sandy′)
- VCA: NP (S\NP)/NP NP
- 'Local' search only – reset one param / input
- Update: Adjust counts/probs., (Re)Set Param to Argmax

Working Memory Cost Metric

After each parse step (Shift, Reduce, Halt):

1. Assign any new Stack entry in the top cell (introduced by Shift or Reduce) a WMC value of 0 (Recency)
2. Increment every Stack cell's WMC value by 1 (Size/Decay)
3. Push the sum of the WMC values of each Stack cell onto the WMC-record (complexity at each step, sum = total complexity)

- Hawkins', Early Immediate Constitutents (EIC)
- Temperley's, Dependency Length Minimization
- Gibson's Processing Costs

Working Memory Cost Metric

After each parse step (Shift, Reduce, Halt):

1. Assign any new Stack entry in the top cell (introduced by Shift or Reduce) a WMC value of 0 (Recency)
2. Increment every Stack cell's WMC value by 1 (Size/Decay)
3. Push the sum of the WMC values of each Stack cell onto the WMC-record (complexity at each step, sum = total complexity)

- Hawkins', Early Immediate Constitutents (EIC)
- Temperley's, Dependency Length Minimization
- Gibson's Processing Costs

Processing Complexity of Constructions / Sentences

- The students who the police who the reporters interviewed arrested laughed (161 C/547 A)
- The students who the reporters interviewed who the police arrested laughed (87)

- daB Peter dem Kunden den Kuhlschrank zu reparieren zu helfen versucht (294)
- daB Peter versucht dem Kunden den Kuhlschrank zu reparieren zu helfen (117)

- He donated the largest single sum ever given by a private individual to the university (C)
- He donated to the university the largest single sum ever given by a private individual (C+20)

- Short $<$ Long (Dep.s & Constit.s) – convergent evolution

Evolutionary Theory and E-Language

1. Linguistic Variation +
2. First Language Learning (Inheritance) +
3. Linguistic Selection / Drift =

Linguistic Evolution

Linguistic variation:

The (E-)language of a speech community is the aggregate output of the distinct I-languages (idiolects) of the changing members of that speech community

Evolutionary Theory and E-Language

1. Linguistic Variation +
2. First Language Learning (Inheritance) +
3. Linguistic Selection / Drift =

Linguistic Evolution

Linguistic variation:

The (E-)language of a speech community is the aggregate output of the distinct I-languages (idiolects) of the changing members of that speech community

The Grammar/Language Set

- 20 P-settings (principles or parameters)
  1. 12 ordering P-settings
  2. 5 category P-settings
  3. 3 rule schemata P-settings
- 8 language 'families' ($\rightsquigarrow$ 270 lgs)
- Sentence Types (3-12)
  1. 3x {s,o1,o2,v} where s,o are 1 word NPs
  2. 5x {s,o1,o2,v} where s,or o is complex
  3. 1x {s,o1,adpos+complex-np, v}
  4. 3x {s,o1,o2,relcl,v}

The Grammar/Language Set

- 20 P-settings (principles or parameters)
  1. 12 ordering P-settings
  2. 5 category P-settings
  3. 3 rule schemata P-settings
- 8 language 'families' ($\leadsto$ 270 lgs)
- Sentence Types (3-12)
  1. 3x {s,o1,o2,v} where s,o are 1 word NPs
  2. 5x {s,o1,o2,v} where s,or o is complex
  3. 1x {s,o1,adpos+complex-np, v}
  4. 3x {s,o1,o2,relcl,v}

The Grammar/Language Set

- 20 P-settings (principles or parameters)
  1. 12 ordering P-settings
  2. 5 category P-settings
  3. 3 rule schemata P-settings
- 8 language 'families' ($\rightsquigarrow$ 270 lgs)
- Sentence Types (3-12)
  1. 3x {s,o1,o2,v} where s,o are 1 word NPs
  2. 5x {s,o1,o2,v} where s,or o is complex
  3. 1x {s,o1,adpos+complex-np, v}
  4. 3x {s,o1,o2,relcl,v}

The Grammar/Language Set

- 20 P-settings (principles or parameters)
  1. 12 ordering P-settings
  2. 5 category P-settings
  3. 3 rule schemata P-settings
- 8 language 'families' ($\rightsquigarrow$ 270 lgs)
- Sentence Types (3-12)
  1. 3x {s,o1,o2,v} where s,o are 1 word NPs
  2. 5x {s,o1,o2,v} where s,or o is complex
  3. 1x {s,o1,adpos+complex-np, v}
  4. 3x {s,o1,o2,relcl,v}

The Grammar/Language Set

- 20 P-settings (principles or parameters)
  1. 12 ordering P-settings
  2. 5 category P-settings
  3. 3 rule schemata P-settings
- 8 language 'families' ($\rightsquigarrow$ 270 lgs)
- Sentence Types (3-12)
  1. 3x {s,o1,o2,v} where s,o are 1 word NPs
  2. 5x {s,o1,o2,v} where s,or o is complex
  3. 1x {s,o1,adpos+complex-np, v}
  4. 3x {s,o1,o2,relcl,v}

The Grammar/Language Set

- 20 P-settings (principles or parameters)
  1. 12 ordering P-settings
  2. 5 category P-settings
  3. 3 rule schemata P-settings
- 8 language 'families' ($\rightsquigarrow$ 270 lgs)
- Sentence Types (3-12)
  1. 3x {s,o1,o2,v} where s,o are 1 word NPs
  2. 5x {s,o1,o2,v} where s,or o is complex
  3. 1x {s,o1,adpos+complex-np, v}
  4. 3x {s,o1,o2,relcl,v}

The Grammar/Language Set

- 20 P-settings (principles or parameters)
  1. 12 ordering P-settings
  2. 5 category P-settings
  3. 3 rule schemata P-settings
- 8 language 'families' ($\rightsquigarrow$ 270 lgs)
- Sentence Types (3-12)
  1. 3x {s,o1,o2,v} where s,o are 1 word NPs
  2. 5x {s,o1,o2,v} where s,or o is complex
  3. 1x {s,o1,adpos+complex-np, v}
  4. 3x {s,o1,o2,relcl,v}

The Grammar/Language Set

- 20 P-settings (principles or parameters)
  1. 12 ordering P-settings
  2. 5 category P-settings
  3. 3 rule schemata P-settings
- 8 language 'families' ($\rightsquigarrow$ 270 lgs)
- Sentence Types (3-12)
  1. 3x {s,o1,o2,v} where s,o are 1 word NPs
  2. 5x {s,o1,o2,v} where s,or o is complex
  3. 1x {s,o1,adpos+complex-np, v}
  4. 3x {s,o1,o2,relcl,v}

The Grammar/Language Set

- 20 P-settings (principles or parameters)
  1. 12 ordering P-settings
  2. 5 category P-settings
  3. 3 rule schemata P-settings
- 8 language 'families' ($\rightsquigarrow$ 270 lgs)
- Sentence Types (3-12)
  1. 3x {s,o1,o2,v} where s,o are 1 word NPs
  2. 5x {s,o1,o2,v} where s,or o is complex
  3. 1x {s,o1,adpos+complex-np, v}
  4. 3x {s,o1,o2,relcl,v}

The Grammar/Language Set

- 20 P-settings (principles or parameters)
  1. 12 ordering P-settings
  2. 5 category P-settings
  3. 3 rule schemata P-settings
- 8 language 'families' ($\leadsto$ 270 lgs)
- Sentence Types (3-12)
  1. 3x {s,o1,o2,v} where s,o are 1 word NPs
  2. 5x {s,o1,o2,v} where s,or o is complex
  3. 1x {s,o1,adpos+complex-np, v}
  4. 3x {s,o1,o2,relcl,v}

Ranking byAverage WMC

1. "English" – SVO, N-r, RelCl-r etc, C, P (30.67)
2. "German" – SOV-v2, N-r, RelCl-r, Scrbl, C, P (30.75)
3. "EngJap" – SVO, N-left, RelCl-left etc, C, P (38.75)
4. "Japanese" – SOV N-left, RelCl-left, scrambling, C, P (40.08)
5. "English" – SVO, N-right, RelCl-right etc, P (61.67)
6. "Japanese" – SOV N-left, RelCl-left etc, C, P (67.83)
7. "English-subset" – SVO, N-simple, RelCl-right, C (61.67)
8. . . .

Population ILM

- Population: $\{LAgt_1, LAgt_2, \ldots LAgt_n\}$
- Language Agent: $(LAgt_i)$

  $< lg^j = LP(UG, fm_k), m_k = Parse(lg^j, f_k),$
  $f_k = Generate(lg^j, m_k), Age(0:9) >$
- Interaction: $(LAgt_i, LAgt_j), i \neq j,$
  $f_k = Generate(lg^i, m_k), m_l = Parse(lg^j, f_k)$
- Interaction Cycle: (mean 30 ints. / LAgt)
  increment Age; $Age(0:3)$ learn; $Age(0:9)$ interact
- Population Initialisation / Replacement: $Age > 9$ replace with $Age = 0 LAgt$, start with mostly adults of various ages, same/diff. P-settings

Population ILM

- Population: $\{LAgt_1, LAgt_2, \ldots LAgt_n\}$
- Language Agent: $(LAgt_i)$

  $< lg^j = LP(UG, fm_k), m_k = Parse(lg^j, f_k),$
  $f_k = Generate(lg^j, m_k), Age(0:9) >$
- Interaction: $(LAgt_i, LAgt_j), i \neq j,$
  $f_k = Generate(lg^i, m_k), m_l = Parse(lg^j, f_k)$
- Interaction Cycle: (mean 30 ints. / LAgt)
  increment Age; $Age(0:3)$ learn; $Age(0:9)$ interact
- Population Initialisation / Replacement: $Age > 9$ replace with $Age = 0 LAgt$, start with mostly adults of various ages, same/diff. P-settings

Population ILM

- Population: $\{LAgt_1, LAgt_2, \ldots LAgt_n\}$
- Language Agent: $(LAgt_i)$

  $< lg^j = LP(UG, fm_k), m_k = Parse(lg^j, f_k),$
  $f_k = Generate(lg^j, m_k), Age(0:9) >$
- Interaction: $(LAgt_i, LAgt_j), i \neq j,$
  $f_k = Generate(lg^i, m_k), m_l = Parse(lg^j, f_k)$
- Interaction Cycle: (mean 30 ints. / LAgt)
  increment Age; $Age(0:3)$ learn; $Age(0:9)$ interact
- Population Initialisation / Replacement: $Age > 9$ replace with $Age = 0 LAgt$, start with mostly adults of various ages, same/diff. P-settings

Population ILM

- Population: $\{LAgt_1, LAgt_2, \ldots LAgt_n\}$
- Language Agent: $(LAgt_i)$

  $< lg^j = LP(UG, fm_k), m_k = Parse(lg^j, f_k),$
  $f_k = Generate(lg^j, m_k), Age(0:9) >$
- Interaction: $(LAgt_i, LAgt_j), i \neq j,$
  $f_k = Generate(lg^i, m_k), m_l = Parse(lg^j, f_k)$
- Interaction Cycle: (mean 30 ints. / LAgt)
  increment *Age*; $Age(0:3)$ learn; $Age(0:9)$ interact
- Population Initialisation / Replacement: $Age > 9$ replace with $Age = 0 LAgt$, start with mostly adults of various ages, same/diff. P-settings

Population ILM

- Population: $\{LAgt_1, LAgt_2, \ldots LAgt_n\}$
- Language Agent: $(LAgt_i)$

  $< lg^j = LP(UG, fm_k), m_k = Parse(lg^j, f_k),$
  $f_k = Generate(lg^j, m_k), Age(0:9) >$
- Interaction: $(LAgt_i, LAgt_j), i \neq j,$
  $f_k = Generate(lg^i, m_k), m_l = Parse(lg^j, f_k)$
- Interaction Cycle: (mean 30 ints. / LAgt)
  increment Age; $Age(0:3)$ learn; $Age(0:9)$ interact
- Population Initialisation / Replacement: $Age > 9$ replace with $Age = 0 LAgt$, start with mostly adults of various ages, same/diff. P-settings

Processibility and Change

Suppose:
$m_i = Parse(lg, f_i)$ fails $\propto$ $WMC(f_i)$ or
$f_i = Generate(lg, m_i)$ is $\propto$ $WMC(f_i)$

$LP(UG, fm_i)$ will be relatively insensitive to higher WML sentences and thus to parameters only manifested in them

VO/OV $\rightarrow$ Pre/Post-Positions:

Kim kissed Sandy in Paris
Kim Sandy Paris+in kissed

Independent parameters during learning
but e.g. WML(OV+Post) < WML(OV+Prep))

Processibility and Change

Suppose:
$m_i = Parse(lg, f_i)$ fails $\propto$ $WMC(f_i)$ or
$f_i = Generate(lg, m_i)$ is $\propto$ $WMC(f_i)$

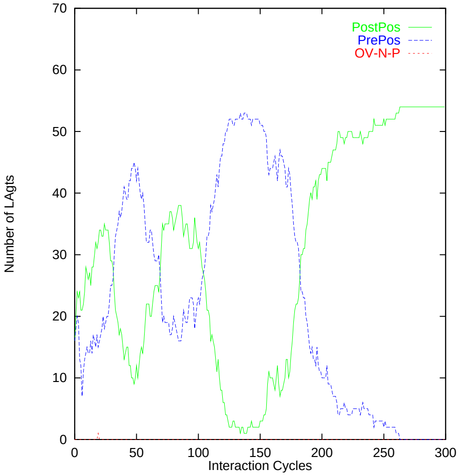
$LP(UG, fm_i)$ will be relatively insensitive to higher WML sentences and thus to parameters only manifested in them

VO/OV $\rightarrow$ Pre/Post-Positions:

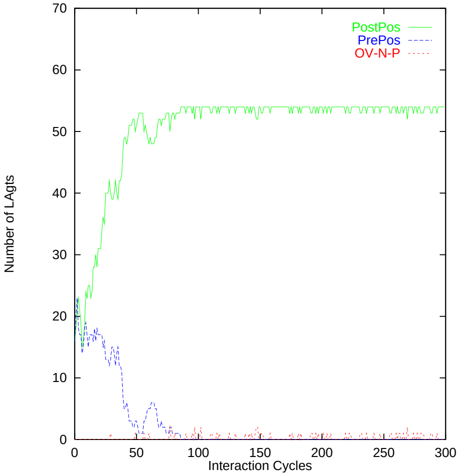
Kim kissed Sandy in Paris
Kim Sandy Paris+in kissed

Independent parameters during learning
but e.g. WML(OV+Post) < WML(OV+Prep))

OV+Prep/Post without processing costs



OV+Prep/Post with processing costs



Zipf's Law & Guirard's/Heap's Law

Straight(ish) lines on log-log plots of freq. vs. rank:

$$c(w) \propto \frac{1}{r(w)^B} \tag{1}$$

$c(w)$ token count of word type w

$r(w)$ rank of word type w in the list of word types sorted in descending order of frequency

$2 > B > 1$, the exponent $=$ slope of the plot

$$V \propto N^A \tag{2}$$

The number of word types V in a text is proportional to the length of that text N

Zipf Curves

Plot of e.g. word frequency against rank deviates from straight line because relative frequency of very common word types closer than the power law predicts, as is relative frequency of very rare words in the tail of the distribution.

of is not half as improbable as the
Many words occur once in the 'long tail'

Power Law (Approximations) Everywhere

- Populations of cities
- Popularity (accesses/links) of web pages
- Relative sizes of earthquakes
- 'Rich get richer' – positive feedback effects
- Dynamical – scale invariance, birth-death processes

Power Law (Approximations) Everywhere

- Populations of cities
- Popularity (accesses/links) of web pages
- Relative sizes of earthquakes
- 'Rich get richer' – positive feedback effects
- Dynamical – scale invariance, birth-death processes

Power Law (Approximations) and Language

- Length / Polysemy vs. freq. / predictability of words
- N-grams of words: bigrams, trigrams,...
- Rules in stochastic grammars: e.g. PCFGs
- Construction type and length
- Word cooccurrence / lexical relations (graphs)

Large numbers of rare types

Probability distribution? (Doubly exponential, Poisson mixtures)

Tail of low counts unreliable – what remains invariant is the shape of the plot not the ranking of types along it or even the set of types:

- egregious, serendipity, globesity (CUP Dicts. On-line)
- Not!, Whatever!

Statistical NLP: Lg model smoothing & adaptation! (samples are not representative)

Large numbers of rare types

Probability distribution? (Doubly exponential, Poisson mixtures)

Tail of low counts unreliable – what remains invariant is the shape of the plot not the ranking of types along it or even the set of types:

- egregious, serendipity, globesity (CUP Dicts. On-line)
- Not!, Whatever!

Statistical NLP: Lg model smoothing & adaptation! (samples are not representative)

Large numbers of rare types

Probability distribution? (Doubly exponential, Poisson mixtures)

Tail of low counts unreliable – what remains invariant is the shape of the plot not the ranking of types along it or even the set of types:

- egregious, serendipity, globesity (CUP Dicts. On-line)
- Not!, Whatever!

Statistical NLP: Lg model smoothing & adaptation! (samples are not representative)

Large numbers of rare types

Probability distribution? (Doubly exponential, Poisson mixtures)

Tail of low counts unreliable – what remains invariant is the shape of the plot not the ranking of types along it or even the set of types:

- egregious, serendipity, globesity (CUP Dicts. On-line)
- Not!, Whatever!

Statistical NLP: Lg model smoothing & adaptation! (samples are not representative)

Word Trends / Volatility

1. Time-ordered corpus of texts (t_i)
2. Continuously compounded return: $r_w(t) = log\frac{f_w(t)}{f_w(t-1)}$
3. Variance / Volatility of return: $std(r_w(t))$
4. Trend of return: $mean(r_w(t)$
5. In grammatical dependency contexts...

neuron(al)/neural in NIPS papers btwn '87-'99 – overall trend flat, highest volatility and trend when modified by noun or adjective – only now useful in field when differentiated: e.g. mirror neuron

'Small World' Graphs (Ferrer-i-Cancho)

1. Growth: at each time step add a node
2. Preferential Attachment: link new node to old nodes with probability proportional to their number of existing links

- Graphs evolve to a scale-invariant organisation
- Power law distribution of nodes by no. of links
- Average path length between nodes is small

Lgs are full of small world graphs: word cooccurrence, dependencies...

'Small World' Graphs (Ferrer-i-Cancho)

1. Growth: at each time step add a node
2. Preferential Attachment: link new node to old nodes with probability proportional to their number of existing links

- Graphs evolve to a scale-invariant organisation
- Power law distribution of nodes by no. of links
- Average path length between nodes is small

Lgs are full of small world graphs: word cooccurrence, dependencies...

'Small World' Graphs (Ferrer-i-Cancho)

1. Growth: at each time step add a node
2. Preferential Attachment: link new node to old nodes with probability proportional to their number of existing links

- Graphs evolve to a scale-invariant organisation
- Power law distribution of nodes by no. of links
- Average path length between nodes is small

Lgs are full of small world graphs: word cooccurrence, dependencies...

Zipfian-ILM Assumptions (Kirby)

Assumptions:

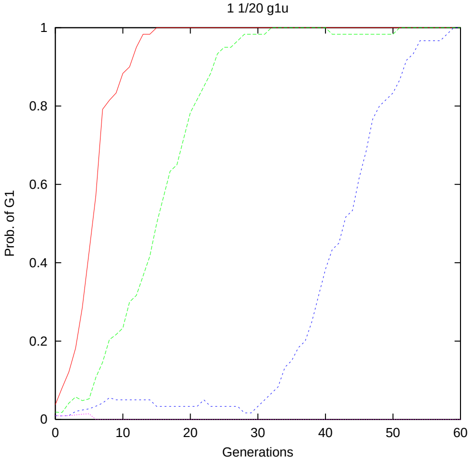
1. an invention strategy for form-meaning pairs,
2. a production bias to express meanings using short forms,
3. a learning bias to learn small grammars and lexicons,
4. a learning period in which not all form-meaning pairs appear
5. and environmental structure which favours some meanings

Zipf-like distributions of words and grammatical rules emerge

S-Curves / Logistic Change

- Logistic / sigmoid is an idealisation (infinite population)
- Kroch used it as a tool to demonstrate a single underlying rate of change in a diverse range of M.Eng. constructions (1 parameter)
- Ellegard's original graphs of constructions are not smooth (finite)
- Emergent from (directed) adaptive change in a finite population of LAgts
- Logistic Map is inherently dynamical (and potentially chaotic)
- A relationship between Zipf-Curves and S-Curves? – they are both strong cues to inherently dynamical (historical) processes, only directed adaptive change results (reliably) in a S-curve

G1 vs. G2 where prior favours G1



Summary

- Variation in E-lg causes drift based on freq-dependent selection until used up
- Adding adaptation to WMC anywhere in LAgts leads to S-curves along typologically plausible lines
- Power laws and S-curves show E-lgs are dynamical systems (not probabilistic generative static stringsets)
- Power laws and s-curves are intuitively related – interderivable?
- Parametric learning provides an account of change in typologically plausible ways when combined with adaptation
- Expressivity and Inductive Bias? Learning Costs?

Reading

Kirby, S. “Spontaneous Evolution of Linguistics Structure: an ILM of the emergence of regularity and irregularity”
Ferrer-i-Cancho, R. “Hubiness, length, crossings and their relationships in dependency trees”
http://www.langev.com/
Hawkins, J. *A performance Theory of Order and Constituency*, CUP, 1994.