COMPARING A LINGUISTIC AND A STOCHASTIC TAGGER
The paper mentions a statistical and a stochastic tagger

Should be a statistical and a linguistic tagger

Terms are used interchangeably in the paper

Stochastic means something relating to conjecture or randomness
What was compared

- The performance of a tagger based on a set of linguistic constraints
- The performance of a tagger based on statistical analysis of text
An updated version of the EngCG tagger – EngCG2

3600 hand coded constraints!

Consists of the following sequentially applied modules:

- Tokenization
- A morphological analysis consisting of:
  - A Lexical component
  - A rule based guesser for unknown words
- Resolution of morphological ambiguities
Operates on a reduced tagset of its own
Tagset is grammatically rather than semantically motivated
Would most likely need to be rewritten to a large degree for other languages
Highly accurate with low ambiguity
A classical trigram-based Hidden Markov Models decoder
Calculates the most likely tag for a word based on the words surrounding it
Creates a reverse suffix tree
Statistical smoothing is performed by reversing up the tree
Employs a more widely used tagset
Can be used on any language provided a sufficiently large corpus exists
Learning curve levels off at around 322,000 words for English
Acquires increased accuracy at the cost of increased ambiguity
To quell criticism of the EngCG tagger

Such as:

- The tagset is overly simplistic
- The accuracy of the EngCG tagger has been overstated
- EngCG trades off high accuracy for high ambiguity
The setup

- A sample of 357,000 words from the Brown corpus
- Tagged by EngCG and human corrected where needed
- A held out benchmark corpus of ~55,000 words from various texts
- Annotated by preprocessor and morphological analyser.
- Fully disambiguated by 2 experts
Each tagger was run on the benchmark corpus
Error rates and ambiguity compared
Also the learning curve for the statistical tagger was measured
Statistical tagger’s learning curve
Statistical analyser’s ambiguity/accuracy tradeoff

Error-rate-ambiguity trade-off

Error rate (%)

Remaining ambiguity (Tags/Word)
The results

- EngCG2 performed radically better
- The statistical tagger’s error rates higher by a factor of 8.6 to 28.0!
- The ambiguity remained very low for very low error rates: 0.10% at 1.070 tags per word
<table>
<thead>
<tr>
<th>Ambiguity (Tags/word)</th>
<th>Error rate (%)</th>
<th>Statistical Tagger (δ)</th>
<th>EngCG (γ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>4.72</td>
<td>4.68</td>
<td></td>
</tr>
<tr>
<td>1.012</td>
<td></td>
<td>4.20</td>
<td></td>
</tr>
<tr>
<td>1.025</td>
<td></td>
<td>3.75</td>
<td></td>
</tr>
<tr>
<td>1.026</td>
<td></td>
<td>(3.72)</td>
<td>0.43</td>
</tr>
<tr>
<td>1.035</td>
<td></td>
<td>(3.48)</td>
<td>0.29</td>
</tr>
<tr>
<td>1.038</td>
<td></td>
<td>3.40</td>
<td></td>
</tr>
<tr>
<td>1.048</td>
<td></td>
<td>(3.20)</td>
<td>0.15</td>
</tr>
<tr>
<td>1.051</td>
<td></td>
<td>3.14</td>
<td></td>
</tr>
<tr>
<td>1.059</td>
<td></td>
<td>(2.99)</td>
<td>0.12</td>
</tr>
<tr>
<td>1.065</td>
<td></td>
<td>2.87</td>
<td></td>
</tr>
<tr>
<td>1.070</td>
<td></td>
<td>(2.80)</td>
<td>0.10</td>
</tr>
<tr>
<td>1.078</td>
<td></td>
<td>2.69</td>
<td></td>
</tr>
<tr>
<td>1.093</td>
<td></td>
<td>2.55</td>
<td></td>
</tr>
</tbody>
</table>
The new version of the EngCG tagger had been created using the Brown corpus as a benchmark.

The EngCG2 tagger had therefore been trained on the benchmark corpus.

The radically better performance required 3600 hand coded rules!