Quantitative Measures of Productivity and their significance
a work-in-progress report (sorry!)

Stefan Evert
Institute of Cognitive Science
University of Osnabrück, Germany
stefan.evert@uos.de

Birmingham, 22 July 2011

Outline

1 Introduction
2 Corpus evidence
3 Productivity measures
4 LNRE models
5 First results
6 Thank you

Introduction

What we want to measure

- Productivity: qualitative vs. quantitative
  - productivity of morphological word-formation rules (e.g. Schultink 1961; Baayen 1992; Evert and Lüdeling 2001)
  - also lexico-grammatical patterns (construction grammar), collocational patterns, word senses, ...

- Vocabulary richness
  - stylometrics & register variation (Baayen 2001, 184-191)
  - authorship attribution (cf. Juola 2006)
  - Zipfian prior for statistical inference (Evert and Pipa 2010)

- Size of the (potential) vocabulary
  - How many words did Shakespeare know? (Efron and Thisted 1976) — And how many typos are there on the Internet?
  - coverage estimation of NLP grammars, dictionaries, ...
  - early indicator for Alzheimer’s disease (Garrard et al. 2005)

Example data

- Bare singulars in English
  - (go) to school, (live) at home, (do) by hand, (come) into effect, (draw) to scale, (fall) in line, (require) by law, ...
  - some authors claim that these are lexicalised exceptions (esp. in German, cf. counter-argument by Kiss 2007)

- Corpus evidence
  - Brown corpus, spoken BNC, written BNC
  - automatic extraction of V + Prep + N sequences
  - only count nouns with ≥ 15% plural occurrences in BNC
  - no manual correction (yet)

- Data extracted with CQP query
  - [class = "VERB"] @[pos = "PR[PF]"
  - [pos = "NN.*"] & [pos = "NN1" & hw = $countable]
  - [pos != "CRD|NN.*"
  - :: match.text_mode = "spoken";
Example data

<table>
<thead>
<tr>
<th>tokens</th>
<th>types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>1,005</td>
</tr>
<tr>
<td>BNC spoken</td>
<td>6,766</td>
</tr>
<tr>
<td>BNC written</td>
<td>85,750</td>
</tr>
</tbody>
</table>

spoken BNC will be used in most of the following examples

Corpus evidence: Zipf ranking

- described by Zipf's law
- popular Zipf-Mandelbrot version (Mandelbrot 1962)

\[ f_k = \frac{C}{(k + b)^a} \]

with "slope" \( a \geq 1 \)

Corpus evidence

- evidence for productivity, type richness and vocabulary size: type-token statistics
- large number of types + many low-frequency types \( \rightarrow \) high degree of productivity
- often shown as Zipf ranking with typical L-shape

Corpus evidence: Zipf ranking

- described by Zipf's law
- popular Zipf-Mandelbrot version (Mandelbrot 1962)

\[ f_k = \frac{C}{(k + b)^a} \]

with "slope" \( a \geq 1 \)

- easily visible as straight line in log-log plot
Corpus evidence: Frequency spectrum

- low-frequency types are better captured by the frequency spectrum
- class size $V_m$ = number of types that occur $m$ times
  - $V_1$ = hapax legomena
  - $V_2$ = dis legomena

Corpus evidence: Vocabulary growth

- vocabulary growth curve shows how number of seen types increases across corpus (+ hapaxes, dis legomena, . . .)
- plot number of seen types $V$ against number of tokens $N$
- slope of VGC = how often new type is encountered

Quantitative measures of productivity

(see Baayen 2001, 24–30)

- Yule (1944) / Simpson (1949)
  \[ K = 10000 \cdot \sum_m m^2 V_m - N \]

- Guiraud (1954)
  \[ R = \frac{V}{\sqrt{N}} \]

- Sichel (1975)
  \[ S = \frac{V_2}{V} \]

- Herdan’s law (1964)
  \[ C = \log V \log N \]

- Baayen’s productivity index (slope of vocabulary growth curve)
  \[ P = \frac{V_1}{N} \]

- TTR = token-type ratio
  \[ TTR = \frac{N}{V} \]

- Zipf-Mandelbrot slope
  \[ a \]

- population size
  \[ S = \lim_{N \to \infty} V(N) \]
Productivity measures for bare singulars in the BNC

<table>
<thead>
<tr>
<th></th>
<th>spoken</th>
<th>written</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>2,039</td>
<td>12,876</td>
</tr>
<tr>
<td>N</td>
<td>6,766</td>
<td>85,750</td>
</tr>
<tr>
<td>K</td>
<td>86.84</td>
<td>28.57</td>
</tr>
<tr>
<td>R</td>
<td>24.79</td>
<td>43.97</td>
</tr>
<tr>
<td>S</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td>C</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>P</td>
<td>0.21</td>
<td>0.08</td>
</tr>
<tr>
<td>TTR</td>
<td>3.32</td>
<td>6.66</td>
</tr>
<tr>
<td>a</td>
<td>1.18</td>
<td>1.27</td>
</tr>
<tr>
<td>pop. S</td>
<td>15,958</td>
<td>36,874</td>
</tr>
</tbody>
</table>

Are these “lexical constants” really constant?

- Comparability (corpus size)
  - do measures depend systematically on corpus size?
- Sampling variation
  - significance tests for differences, confidence intervals
- Non-randomness (Baroni and Evert 2005, 2007)
- Manual data correction
  - not feasible for large samples, e.g. 85,750 types in BNC
- Interpretation of productivity measures
  - productivity vs. vocabulary richness vs. size of vocabulary
  - does any measure match our intuition of productivity?

Extrapolation with LNRE models

- direct comparison of written vs. spoken BNC not possible
- productivity measures need to be perfectly size-invariant or sample size has to be adjusted (to larger sample)
- use statistical LNRE models (Khmaladze 1987; Baayen 2001; Evert 2004a,b) to extrapolate vocabulary growth
Extrapolation with LNRE models

- Direct comparison of written vs. spoken BNC not possible
- Productivity measures need to be perfectly size-invariant or sample size has to be adjusted (to larger sample)
- Use statistical LNRE models (Khmaladze 1987; Baayen 2001; Evert 2004a,b) to extrapolate vocabulary growth
- Extrapolation of frequency spectrum also possible

LNRE models as a methodological research tool

- LNRE models can also help us to learn more about the properties of productivity measures
- Separate variability of measures into:
  - Size dependency (expected spectrum for different N)
  - Sampling variation (parametric bootstrap samples) under controlled conditions
- Quantify sampling variation → significance tests, etc.
- Mature & user-friendly implementation for Gnu R in the zipfR package (Evert and Baroni 2007)

First results

Which measures are size-invariant?

Expected frequency spectrum factors out effects of sampling variation

How much are measures affected by sampling variation?

Are the differences between spoken and written BNC significant?
How much are measures affected by sampling variation?
Zipf slope and population size estimated from trained LNRE model

Sample size matters!
Brown corpus is too small for reliable LNRE parameter estimation

Sample size matters!
other productivity measures seem to be more robust

Thank you
There's much work to be done, of course!
Talk about interpretation of measures in the coffee break?
References I


References II


Evert, Stefan and Pipa, Gordon (2010). Probability estimation of rare events in linguistics and computational neuroscience. Presentation at the KogWis 2010 Conference, Potsdam, Germany.


References III


