Jurgis Pakerys (Vilnius University)
Measuring morphological productivity

jurgis.pakerys@flf.vu.lt
1. Productivity and frequency

2. Measuring productivity
   2.1. Sources of measurements
   2.2. Realized productivity
   2.3. Hapax-based measures of productivity
      2.3.1. Expanding productivity
      2.3.2. Potential productivity

4. Summary

5. References
1. Productivity and frequency

Morphological processes related to **lexemes**:

- Composition
- Derivation
- Assignment to **inflectional classes**
  (= declinations, conjugations)
- Grammatical forms
1. Productivity and frequency

Frequency vs. productivity

- **Frequent** = abundant = affects many members
- **Productive** = alive = attracts/produces many NEW members
1. Productivity and frequency

Understanding frequency

- **Token** frequency = number of times a lexeme occurs in the corpus

- **Type** frequency = number of times a morphological process is found in all lexemes of the corpus
1. Productivity and frequency

**Type vs. token**, artificial example

- **Token** frequency of *mängi-mine* is 567 = various forms of this N occur 567 times in a given corpus

- **Type** frequency of *-mine* is 14232 = suffix *-mine* is found 14232 times in the list of lexemes (not their forms!) of a given corpus
1. Productivity and frequency

Combinations of frequency and productivity

1. Frequent and Productive
   • High type frequency
   • Attracts new members

2. Frequent and Non-Productive
   • High type frequency
   • Does not attract new members
1. Productivity and frequency

Combinations of frequency and productivity

3. Productive and Non-Frequent
   • Attracts new members
   • Low type frequency

4. Non-productive and Non-Frequent
   • Does not attract new members
   • Low type frequency
2. Measuring productivity

2.1. Sources of measurements

- Dictionaries
- Corpora
- Questionnaires, tests
  - Open-ended coinage tests, judgment tasks
    (see, for example, Bolozyk 1999)
2.2. Realized productivity

- Number of the members of the morphological process in a dictionary / corpus
- Realized productivity, extent of use (Baayen 2009: 904)
- Frequency = / ≠ productivity
- Neologisms!
2.2. Realized productivity

Doing it:

• Get a traditional **dictionary** or a **list** of all lemmas of the corpus

• **Filter** by affix (+ any additional parameters available); what about compounds?
2.2. Realized productivity

- **Clean** the data manually (synchronously non-derived items, non-affixes, etc.)
- **Delete** inner derivational cycles (optional), cf. English:
  - *decompos-able* $<$ *de-compose* $<$ *compose*
  - *de-* should count as a derivational affix in decomposable
2.2. Realized productivity

Example (Gaeta & Ricca 2006)

• Corpus study (*La Stampa*, 1996-98, 75M)
• **Counting types, \( V(N) \), vertical axis**
• Counting tokens, \( N \), horizontal axis

1. *-mente*: adverb
2. *-mento, -(t)ura, -nza*: action noun
**N**: Token number of the suffix

Fig. 47.1: Vocabulary growth curve $V(N)$ for four Italian derivational suffixes (from Gaeta and Ricca 2006: 58)
2.2. Realized productivity

Criticizing it:

• Realized productivity shows how productive a morphological process was in the PAST

• What processes are attracting new members NOW? What about the FUTURE?
2.3. Hapax-based measures of productivity

- *Hapax (legomenon)*
- Attested only once in a corpus
- Sometimes ignored as rubbish (numbers, typos, crazy character sequences, etc.)
2.3. Hapax-based measures of productivity

- Correlation between hapaxes and new formations/new borrowings

- Do not just believe it, let’s think: why new words are rare?
2.3. Hapax-based measures of productivity

• Note: not all hapaxes are new words, but it is fine, they are just a good statistical indicator! (cf. Baayen 2009: 906)

• Size matters: the bigger, the better (?) (see Baayen 1993: 189, 2009: 905)
2.3. Hapax-based measures of productivity

Two hapax-based measures

- **Expanding** productivity
- **Potential** productivity

- See Baayen 1993, 2009: 905-907
2.3.1. Expanding productivity

- $V(1,N)$, the number of (derivationally transparent) hapaxes with the affix $X$
- $V(1)$, the total number of hapaxes of the corpus

$$P^* = \frac{V(1,N)}{V(1)}$$

- $P^*$ shows the market share of the affix in the market of hapaxes (= possibly new words)

Baayen 2008: 902, 905
2.3.1. Expanding productivity

Doing it:

• Get the **list of hapaxes** of a given corpus (DIY or ask for help)

• A lemmatized list of hapaxes helps a lot for a language like Estonian

• **Filter the items** you are interested in (according to the affixes, etc.)

• **Manually** clean the lists (see above on realized productivity)
2.3.1. Expanding productivity

- **Count** $P^*$ values
- **Rank** the morphological processes (affixes, etc.) according to $P^*$
- **Q**: is division by the total number of hapaxes of the corpus necessary?
2.3.1. Expanding productivity

Criticizing it:

• Some processes (affixes, etc.) get extremely high numbers of hapaxes, but they do not seem to be as productive

• Example: Italian deverbal agent suffix -(t)ore (male/generic) has 2x more hapaxes than -trice (female) (Gaeta & Ricca 2006: 73-74)

• Not fair!
2.3.1. Expanding productivity

- Variable corpus approach (Gaeta & Ricca 2006)
- Count hapaxes for equal numbers of tokens of a given process
- For this, the sizes of the subcorpora will be different (= variable corpus)

- Weakness: some affixes do not reach the token frequency needed (then: binominal interpolation, extrapolation)
2.3.1. Expanding productivity

- P* and inflection class (IC) productivity?
- Wurzel 1989: 149 on new formations / loans as indicators of productive ICs

- See esp. Gaeta 2009 on using variable corpus approach to measure inflectional morphology
2.3.2. Potential productivity

- $V(1,N)$, the number of hapaxes with the affix $X$
- $N$, the number of forms of lexemes with the affix $X$ (tokens, lexeme frequency)

$$P = \frac{V(1,N)}{N}$$
2.3.2. Potential productivity

• Higher value of P:
  – the forms of lexemes with the affix X are (still) comparatively rare
  – the affix X has the potential to get a larger share of the onomasiological market (Baayen 2008: 902, 906)

• Alternative: variable corpus approach (count P for equal numbers of tokens of a given affix)
2.3.2. Potential productivity

- Example, Dutch (Baayen 2008: 905-907)
- -ster (deverbal agent, female)
- ver- (verbal prefix)
- -ster should be more productive (intuitively)

- Types (42M corpus): 370 (-ster) vs. 985 (ver-)
- Hapaxes: 161 (-ster) vs. 274 (ver-)
- Potential prod.: 0.031 (-ster) vs. 0.001 (ver-)
2.3.2. Potential productivity

Doing it:

• Get the **list of lexemes with token frequency data**, filter the relevant ones, clean the list manually, count the total token frequency

• Get the **list of hapaxes** (filter the first list, frequency = 1), filter the relevant items, clean the list manually

• **Count P value**, rank the affixes according to it
Summary

• **Realized** productivity

• Hapax-based measures
  – **Expanding** productivity
    (hapaxes with affix $X$ : all hapaxes)
  – **Potential** productivity
    (hapaxes with affix $X$ : tokens with affix $X$)

• Variable corpus approach
References and further reading

• **Website of R. H. Baayen**: http://www.sfs.uni-tuebingen.de/~hbaayen/


References and further reading


References and further reading
