Motivation: a little lexical statistics

- how probable is it that you find a new word after having sampled a certain amount of text?
- this can be computed from the distribution of words in a text (Baayen 1992, 2001)

Aside: counting words

- decide whether you want to count graphemic words or words according to some theory (pigeon hole vs ger.: Postfach)
- token: a word at a certain position in a text
- type: a unique form
  - word-form type
  - lemma type
- type-token statistics

Aside: counting words

A journey of a thousand meals begins with a single bite. My choucroute obsession had taken hold with the first version I had tasted, a recipe of Julia Child’s. It was sweet with chicken broth, onions, and carrots, aromatic with cloves and juniper berries. The meats were roast pork and sausages, bacon and ham, all luxurious and familiar. I had relished every morsel and imagined that a thoroughly authentic choucroute would be the same, only more so. And thus my quest began.

(Jeffrey Steingarten, True Choucroute, from The Man who Ate Everything, 1997, New York)
Aside: counting words

A journey of a thousand meals begins with a single bite. My choucroute obsession had taken hold with the first version I had tasted, a recipe of Julia Child’s. It was sweet with chicken broth, onions, and carrots, aromatic with cloves and juniper berries. The meats were roast pork and sausages, but two word forms I found curious and familiar. I had relished every one of the lemma BEGIN that a thoroughly authentic choucroute would be the same, only more so. And thus my quest began.

(Jeffrey Steingarten, True Choucroute, from The Man who Ate Everything, 1997, New York)

what distributions tell us

rough idea:
- if you have a finite number of words in your language, chances are, you will have encountered all of them more than once after you have looked at a text that is large enough
- if you have an infinite number of words you will have encountered many of them only once (and many not at all, of course)
- so the number of rare words might tell you something about the number of words in your language

N = tokens, V = types:
in the distribution labeled ‘productive’ the number of types continues to grow as new tokens are sampled

calculating a type-token ratio

- count all types in the text
- count how often each type occurs (a type-token ratio)
- for the German Stuttgarter-Zeitung corpus (36 m word newspaper text)

lemma types from STZ-corpus

<table>
<thead>
<tr>
<th>type</th>
<th>frequency</th>
<th>type</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>3,571,573</td>
<td>Zytomegalievirus</td>
<td>1</td>
</tr>
<tr>
<td>.</td>
<td>1,848,517</td>
<td>Zytori</td>
<td>1</td>
</tr>
<tr>
<td>.</td>
<td>1,605,763</td>
<td>Zytori</td>
<td>1</td>
</tr>
<tr>
<td>ein</td>
<td>710,719</td>
<td>Zytos</td>
<td>1</td>
</tr>
<tr>
<td>und ‘and’</td>
<td>708,531</td>
<td>zytotoxische</td>
<td>1</td>
</tr>
<tr>
<td>in ’in’</td>
<td>613,876</td>
<td>Zywietz</td>
<td>1</td>
</tr>
<tr>
<td>PPER</td>
<td>536,174</td>
<td>Zyzik</td>
<td>1</td>
</tr>
<tr>
<td>sein ‘to be’</td>
<td>534,056</td>
<td>ZZ-Top-Hit</td>
<td>1</td>
</tr>
<tr>
<td>*</td>
<td>408,708</td>
<td>ZZ-Top-Käfer-Nachbau</td>
<td>1</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>ZZF-Information</td>
<td>1</td>
</tr>
</tbody>
</table>

calculating a frequency distribution

- count how often each frequency occurs (the frequency of frequencies)
### Calculating a Frequency Distribution

<table>
<thead>
<tr>
<th>freq</th>
<th>freq of freq</th>
<th>freq</th>
<th>freq of freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>404,579</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>96,981</td>
<td>708,531</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>43,357</td>
<td>710,719</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>26,159</td>
<td>1,605,763</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>17,559</td>
<td>1,848,517</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>3,571,573</td>
<td>1</td>
</tr>
</tbody>
</table>

That means:

- more than half of all lemma types in the corpus occur only once (hapax legomena)
- which is evidence that we have not come near sampling all the available words
- similar measures are found for texts of any size (Zipf's law, LNRE distribution (Baayen 2001))
- for all intents and purposes we can assume an infinite number of words

### Kinds of Rare Words

- names (*Zywietz, Zyzik*)
- misspellings
- words from a different genre (*zytotoxisch*)
- words from a different language

In the wide valley far below I could make out a tiny farmer on a tiny tractor lugging a tiny wagon bursting with *quintal d’Alsace*, huge white cabbages that would soon be finely shredded, layered with salt and juniper berries, and fermented into choucroute which is French for “sauerkraut” which is German for “bitter herb”.

(*Jeffrey Steingarten, True Choucroute, from The Man who Ate Everything, 1997, New York*)

### Kinds of Rare Words

- names (*Zywietz, Zyzik*)
- misspellings
- words from a different genre (*zytotoxisch*)
- words from a different language
- words formed by creative processes:
  - *(alleinunterhaltsam, kinobetriebsam)*

This presupposes a morphological theory!
Kinds of rare words

- names
- misspellings
- words from a different genre
- words from a different language
- words formed by creative processes:
- neologisms formed according to morphological regularities
  \(\rightarrow\) morphological productivity

Schultink’s definition

“We see productivity as a morphological phenomenon as the possibility for language users to coin unintentionally an in principle unlimited number of new formations, by using the morphological procedure that lies behind the form-meaning correspondence of some known words.”
Schultink (1961, 113)

Morphological Productivity

No language has a finite vocabulary:
- new (basic) entries through borrowing, creativity, mechanisms like clipping, blending, etc.
- new complex entries through word formation mechanisms: compounding, derivation, conversion, etc.

Productivity: attempts of definition 1

„We see productivity as a morphological phenomenon as the possibility for language users to coin unintentionally an in principle unlimited number of new formations, by using the morphological procedure that lies behind the form-meaning correspondence of some known words.” (Schultink 1961,113, our translation)

Qualitative aspects of productivity

- Goal: an intensional description of the possible bases for a given word formation rule.
- Bases can be constrained
  - morphologically
  - phonologically
  - syntactically
  - semantically
  - pragmatically
- In addition, there can be language-specific general rules on the combination of affixes.

Basic word formation

affixes select for their stems; restrictions involve all linguistic levels
- part-of-speech: -bar attaches to verbs: ess+bar "edible", verwertbar "usable"
- argument structure: -bar attaches to transitive verbs: *schlafbar from intr. schlafen "to sleep"
- morphological: Ge- -e does not attach to complex stems: *Ge+ver+kau+f+e
  -ler attaches to abbreviations, but -lich does not: SPDler "sb. belonging to the SPD", *SPDlich
Basic word formation VII

- phonology: -ei attaches to words that end in an unstressed syllable, otherwise the allomorph -erei is used
- origin (etymology): -bar attaches to native and neoclassical elements, -abel attaches only to neoclassical elements
  annehmbar, akzeptierbar "acceptable"
  akzeptabel, *annehmabel

Basic word formation

- semantics/concept
  -fach attaches to numbers: dreifach "three times", tausendfach "a thousand times"
- Ge- -e does not attach to stative verbs:
  *Gewisse, *Gekenne

Basic word formation

- ideally, every element would be marked in the lexicon for all of these properties so that a word formation component can refer to them when modelling the restrictions

Example 1: suffix -lich I

- information about the suffix itself
  - produces: adjectives
  - origin: native
  - semantic functions: relative, ornative, ...
  - types/tokens produced (in 200 m word form corpus): 873/1473983
  - hapax legomena: 169
  - phonology: not stress-influencing, final devoicing

Example 1: suffix -lich II

- from corpus data & descriptive word-formation literature: find the different patterns for -lich
  - -lich attaches to
    - nouns
    - adjectives
    - verbs
  - for each pattern: find the restrictions on the base and the properties of the pattern
Example 1: suffix -lich III

- pattern: -lich attaches to nouns that denote a profession or human property: bischoflich, ärztlich, amszärztl, richterlich, großmütterlich, studienrätlich, feindlich, fraulich...
- base
  - origin: native (no neoclassical or English or other foreign bases found)
  - morphology: no abbreviations, complex bases possible (derivations, compounds)
- pattern
  - meaning: like a N, from a N
  - use: attributive, adverbial, predicative

Example 1: suffix -lich IV

- pattern: -lich attaches to adjectives: grünlich, säuertich, bräunlich, ...
- base
  - origin: native
  - phonology: monosyllabic, bisyllabic
- pattern
  - meaning: a little Adj
  - use: attributive, adverbial (?), predicative

Example 1: suffix -lich V

- other patterns for -lich
- with 'non-human' nouns: pflanzlich, herbstlich, ...
- with temporal nouns: monatlich, stündig, täglich, ...
- with verbs: beachtlich, bezwinglich,...

Example 1: suffix -lich VI

- lists of 'semi-regular' forms (lexicalized)
- phonological irregularity: flehentlich, versehentlich, wöchentlich, weinerlich, ...
- semantic irregularity: rundlich, kürzlich, freundlich, persönlich, tröstlich, ...
- list of unanalysable forms: freilich, möglicher, redlich, ...

Example II: adjective compounds

- patterns for Adj + Adj (all productive)
  - simplex + simplex: hellblau, frühweise, ...
    - determinative compound, stress on non-head, non-recursive
  - simplex + derivation: giftiggelb, ...
    - determinative compound, stress on non-head, non-recursive
  - ...
  - simplex + simplex: graublauf, sauerscharf, ...
    - copulative compound, stress on non-head + head, recursive

Quantitative notions of productivity:

- degrees of productivity:
  - "quasi-productive, marginally productive, semi-productive, fully productive, most productive, quite productive, immensely productive, very productive" (Plag 1999, 12)
Degrees of productivity II

- How can we measure the degree of productivity for a word formation rule?
- Is 'degree of productivity' an intrinsic property of a rule?

Degrees of productivity III

'The degree of productivity of a WF-rule can be seen as inversely proportional to the amount of competence restrictions on that WF-rule.' (Booij 1977, 5)

Degrees of productivity IV

'There is a simple way to take such restrictions into account: we count up the number of words which we feel could occur as the output of a given WFR (which we can do by counting the number of possible bases for that rule), count up the number of actually occurring words formed by that rule, take a ratio of the two, and compare this with the same ratio for another WFR. In fact, by this method we could arrive at a simple index of productivity for every WFR: the ratio of possible to actual words.' (Aronoff 1976, 36)

Aronoff's measure

- actual words = attested words
- possible words = unattested words

Attested where?

What can a ratio of actual to possible words tell us?

Productive rules vs. unproductive rules

- productive rules produce indefinitely many words
- unproductive rules produce a finite number of words

How can we observe productivity?

- morphological productivity is a psycholinguistic (cognitive) notion
- we can only observe reflexes of productivity
- reflexes can be analysed qualitatively and quantitatively
- this analysis enables us to infer (some) properties of morphological productivity
Measuring morphological productivity I

`Any measure of morphological productivity that is of linguistic interest [...] should provide a ranking of word-formation processes that is in general correspondence with a ranking based on linguistic intuitions [...] should express ‘the statistically determinable readiness with which an element enters into new combinations. [...]`

Taking into account those formations which are characterized by formally or semantically idiosyncratic properties should have the effect of lowering the value of the productivity measure. [...] should shed light on the empirical fact that productivity cannot be simply measured in terms of type frequencies.’ (Baayen 1992, 110f)

Dictionary data

- no frequency information
- often no information on date of first occurrence (but see OED, which lists, for example the twentieth century neologisms)
- contains obsolete words
- most transparent words are not listed

Corpus data

- frequency information
- date of first occurrence not given
- corpus size
- representativeness
- corpus quality

Procedure

- count all occurrences of words (types vs. tokens) formed by a given word-formation process in a corpus
- use statistical methods
  - Zipf ranking
  - frequency spectrum
  - vocabulary growth curve (VGC)

Baayen’s measure of productivity: growth rate

\[ P = \frac{\text{formel hapax durch lemma}}{\text{tokens (V1 durch N)}} \]

\[ P \text{ predicts at what rate new types, types that are not represented in that sample, will appear when we decide to enlarge the sample.' (Baayen & Lieber 1991, 811) \]

What does this tell us about productivity?
Global productivity

- hier Formel: global productivity I ist gleich population size S durch Anzahl der attested types V
- population size S: limes N bis unendlich V(N)
- for productive processes:
  - limes N bis unendlich V1/V ungleich 0 (Proportion der Hapaxe an den Typen größer 0)

P for some word formation processes

- data from a tagged and lemmatized newspaper corpus (Stuttgarter Zeitung)
- find lemmas with the corpus query processor (cqp):
  - [lemma = ".*sam" & pos = "ADJ."];

Aside: CQP data I

- CQP data is in the kwic (keyword in context) format
- 101224: obigt je ein Drittel des <gemeinsamen> Gottesdienstes ausrichte
- 101251: esionen in diesem Moment <gemeinsam> das Grab bewachten , war
- 101385: ei nicht so sehr Ausdruck <gemeinsamen> Schicksals gegenüber der
- 101985: dacht zu haben . Aoki war <gemeinsam> mit über 600 Gästen am 1
- 104410: tte " ( Biermann ) . Die <grausame> Wahrheit dieses Jahrhund
- 105575: ung und halben Heldentums <seltsam> brüchig scheinen . Der l
- 107036: tik betrieben und auf die <gewaltame> Intervention in Polen ge
- 107792: chischen Genossen darauf <aufmerksam> gemacht , daß sie in ers

Aside: CQP data II

One can also group lemma types and get the token count:

<table>
<thead>
<tr>
<th>Lemma</th>
<th>Token Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>gemeinsam</td>
<td>19961</td>
</tr>
<tr>
<td>langsam</td>
<td>4520</td>
</tr>
<tr>
<td>wirksam</td>
<td>2437</td>
</tr>
<tr>
<td>aufmerksam</td>
<td>2234</td>
</tr>
<tr>
<td>gleichsam</td>
<td>1793</td>
</tr>
<tr>
<td>seltsam</td>
<td>1427</td>
</tr>
<tr>
<td>mühsam</td>
<td>1426</td>
</tr>
<tr>
<td>einsam</td>
<td>1335</td>
</tr>
<tr>
<td>bedeutsam</td>
<td>976</td>
</tr>
<tr>
<td>gewatsam</td>
<td>870</td>
</tr>
<tr>
<td>spansam</td>
<td>770</td>
</tr>
</tbody>
</table>

Calculating P

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>N1</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ös</td>
<td>4383</td>
<td>70</td>
<td>0.0160</td>
</tr>
<tr>
<td>-sam</td>
<td>22667</td>
<td>78</td>
<td>0.0034</td>
</tr>
<tr>
<td>-bar</td>
<td>37783</td>
<td>324</td>
<td>0.0086</td>
</tr>
</tbody>
</table>
Are these values intuitive?  
What do they tell us about productivity?

We should perhaps look at the development of P.

Development of P: theory

- if P is an intrinsic value of productivity that 'belongs' to a word formation process it should be static throughout a corpus
Development of $P$: data

- $P$ changes throughout the corpus (the smoother shows the trend but even that changes)
- the value of $P$ becomes smaller
- This is why $P$ is called local productivity: $P$ shows the productivity after a certain amount of text.

Comparing $P$-values

- because of the changes, $P$-values of different word-formation processes cannot be compared

Why do the values change so much?
We have to take a closer look at the data: compare Vocabulary Growth Curves.
Vocabulary growth curves

- they all look productive
- that is not intuitive
- we have to look at the data

Looking at the data I

Looking at the data II
Looking at the data III

Looking at the data III

graziös           44
maliziös          42
amounös           41
schnupös          41
bös               40
wös               1
populär-pompös    1
erwünschtreligiös 1
radikal-religiös  1
fundamentalistisch-religiös 1
melodios-kapriziös 1
aristokratisch-preziös 1

Corpus problems

- spelling mistakes
- repetitivity: repetition of articles, standardized text
- kind of corpus
- size

Pre-processing problems

- tokenizing errors
  
  habe.Aufmerksam

- lemmatization errors
  
  Erstechen, Ausbleichen, Jös, Rös, ...

Linguistic problems I

- structure:
  
  (A(BC)): no new hapax for process C
  ((AB)C): hapax for process C

  Nachmittagschläfchen, Schokokalenderchen, Koalitionspartnerchen, Erdmännchen, Topfpflänzchen, ...

Linguistic problems II

- items that end in (or begin with) a sequence of letters that is identical to the affix we are looking at but are no derivations

  Zeichen, Jochen, München, Erstechen, Groschen, Stachetrochen, bös

Linguistic problems III

- creativity

  kinobetriebsam, alleinunterhaltsam, unkaputtbar

- We need a precise description of the process we are investigating!
Linguistic problems IV

- What should we do with affixes that are not semantically transparent (anymore)? Are they still morphologically recognizable? Can they be a factor in determining the productivity of an affix?

Mädchen, Beffchen, Eichhörnchen, Ohrläppchen, Schnäppchen, Spirenzchen, Kinkerlitzchen

Linguistic problems V

- What do we do with morphologically complex items that were formed by a process that is now unproductive?

dienstbar, jagdbar, fruchtbar, furchtbar
Manually corrected VGCs

- the curves are shorter because the number of considered types is lower – we threw out the mistakes
- the curves are more intuitive: -ős and -sam now look unproductive

P for manually corrected data

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>N₁</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>-ős</td>
<td>3404</td>
<td>5</td>
<td>0.0025</td>
</tr>
<tr>
<td>-sam</td>
<td>22654</td>
<td>5</td>
<td>0.0002</td>
</tr>
<tr>
<td>-bar</td>
<td>35562</td>
<td>189</td>
<td>0.0053</td>
</tr>
</tbody>
</table>
Comparing VGCs

- they differ in length
- we know that the productivity index is dependent on the size of the corpus
- can we compare the P values?
- We need a way of predicting how these curves would continue with growing N: this is called extrapolation, and we need statistical models to extrapolate

References

- Chomsky, Noam (1957) Syntactic Structures. Mouton, den Haag